

DELHI COLLEGE OF ENGINEERING



सत्यमेव जयते

LIBRARY



Class No.

Book No.

Accession No.



Automatica

The Journal of IFAC the International Federation of Automatic Control

Volume 27 Number 1

January 1991

CONTENTS

	1	Call for Papers
		PAPERS
D. P. Bertsekas and N. Tsitsiklis	3	Some Aspects of Parallel and Distributed Iterative Algorithms: A Survey
R. O. LaMaire, L. Velavan, M. Athans and G. Stein	23	A Frequency domain Estimator for Use in Adaptive Control Systems
T. T. Tuij and J. B. Moore	39	Enhancement of Fixed Controllers via Adaptive Q Disturbance Estimate Feedback
H. Demircioğlu and P. T. Gawthrop	55	Continuous time Generalized Predictive Control (CGPC)
J. Ackermann, D. Keesbauer and R. Münch	75	Robust Gamma stability Analysis in a Plant Parameter Space
C. I. Byrnes and A. Isidori	87	On the Attitude Stabilization of Rigid Spacecraft
G. Johansen and J. L. Alty	97	Knowledge Engineering for Industrial Expert Systems
L. Wehenkel and M. Pavella	115	Decision Trees and Transient Stability of Electric Power Systems
		BRIEF PAPERS
N. L. Segall, J. F. MacGregor and J. D. Wright	135	One step Optimal Saturation Correction
A. İftar and U. Özgüner	141	Modeling of Uncertain Dynamics for Robust Controller Design in State Space
A. Tesi and A. Vicino	147	Robust Absolute Stability of Lur'e Control Systems in Parameter Space
F. Giri, M. M. Saad, J. M. Dion and L. Dugard	153	On the Robustness of Discrete time Indirect Adaptive (Linear) Controllers
K. Gu, Y. H. Chen, M. A. Zohdy and N. K. Loh	161	Quadratic Stabilizability of Uncertain Systems: A Two Level Optimization Setup
D. C. Hyland and E. G. Collins, Jr.	167	Some Majorant Robustness Results for Discrete time Systems
C. C. H. Ma	173	Rapid Tracking of Complex Trajectories in Short-duration Processes
Y. C. Juan and P. T. Kabamba	177	Optimal Hold Functions for Sampled Data Regulation
T. Söderström, P. Stoica and B. Friedlander	183	An Indirect Prediction Error Method for System Identification
S. L. Campbell	189	Comments on 2-D Descriptor Systems
D. Mustafa, K. Glover and D. J. N. Limebeer	193	Solutions to the H^∞ General Distance Problem which Minimize an Entropy Integral

Continued on outside back cover**Pergamon Press**

OXFORD

NEW YORK

BEIJING

FRANKFURT

SÃO PAULO

SEOUL

SYDNEY

TOKYO

automatica

The IFAC Journal

By an agreement between IFAC and Pergamon Press plc (the official IFAC publisher), AUTOMATICA is the official Journal of IFAC, the International Federation of Automatic Control

IFAC Council

President B. D. O. Anderson
President elect S. J. Kahne
Vice president L. Ljung
Vice president Y. Z. Li
Immediate Past President B. Tamim
Treasurer M. Mansour
Ordinary Members
J. Ackermann
A. van Cauwenberghie
E. J. Davison
A. Ichikawa
V. Kucera
P. M. Larsen
A. Titli
J. D. N. van Wyk

Technical Board

Chairman L. Ljung
Vice chairman
E. J. Davison
I. Keviczky
P. M. Larsen
M. G. Rodd
P. Urosien

Editorial Board

Chairman G. S. Axelby
Vice chairman
H. A. Spang III
H. Kwakernaak
Chairman Publication Committee G. Guardabassi
Chairman of IFAC Publications
Managing Board M. Thoma
Members
K. J. Åström
P. M. Larsen
W. S. Levine
P. C. Parks
M. Rodd Technical Board Liaison
A. P. Sage
P. T. Shepherd
R. E. Strang

Executive Board

Chairman Y. Z. Li
Immediate Past President B. Tamim
Treasurer M. Mansour
Secretary G. Housney
Chairman Policy Committee P. M. Larsen
Chairman Publication Committee G. Guardabassi
President elect S. J. Kahne

IFAC Secretariat

G. Housney
B. Aumann
E. Rordus
Schlossplatz 12
2301 Laxenburg
Austria

Parameter Estimation and Adaptive Control

Patrick C. Parks
Mathematics Group, School of Defence Management
Royal Military College of Science
Shrivenham, Swindon, SN6 8LA, U.K.

Large-scale Systems, Management and Decision Sciences

Andrew P. Sage
George Mason University
4400 University Drive
Fairfax, VA 22030, U.S.A.

Survey Papers

Karl J. Åström
Division of Automatic Control
Lund Institute of Technology
S-221 00 Lund, Sweden

Technical Communications and Correspondence, Rapid Publications

William S. Levine
Dept of Electrical Engineering
University of Maryland, MD 20742, U.S.A.

Book Reviews

Peter Martin Larsen
Electrical Power Engineering Dept
Bldg 325, Technical University of Denmark
2800 Lyngby, Denmark

Authors should send five copies of manuscripts for publication to appropriate Editor and one copy to the Editor-in-Chief with copy of letter to Editor.

Associate Editors

- Y. Akaiwa Georgia Institute of Technology, Atlanta, U.S.A.
K. E. Åzmi Lund Institute of Technology, Lund, Sweden
A. Bagchi Univ. of Toronto, The Netherlands
I. Basar Univ. of Illinois, U.S.A.
R. Bitsum Australian National Univ., Canberra, Australia
R. Canales Univ. Tecnológico de Chihuahua, U.T. Chihuahua del Estado, Chihuahua, Mexico
A. van Cauwenberghie Rijksuniversiteit Gent, Ghent, Belgium
J. H. Chow G.E. FUSEL, Schenectady, NY, U.S.A.
D. W. Clarke Univ. of Oxford, Oxford, U.K.
R. Corbett Univ. of Groningen, The Netherlands
P. Dorato Univ. of New Mexico, Albuquerque, NM, U.S.A.
P. M. G. Taming Pontificia Univ. Católica, Rio de Janeiro, Brazil
B. Fordham New Jersey Institute of Technology, Newark, NJ, U.S.A.
P. J. Gawthrop The University, Glasgow, U.K.
G. Guardabassi Politecnico di Milano, Milano, Italy
Y. Y. Haimas Univ. of Virginia, VA, U.S.A.
C. C. Hsing Nat. Univ. of Singapore, Singapore
M. Ikeda Kyoto University, Japan
M. Ismail Technische Hochschule Darmstadt, F.R.G.
A. Isidori Univ. di Roma, La Sapienza, Rome, Italy
M. Johnson Univ. of Strathclyde, Glasgow, U.K.
S. J. Kahne The MITRE Corp., McLean, VA, U.S.A.
This staff of Associate Editors is being expanded to include representatives from various countries throughout the world.
I. Keviczky Computers and Automation Institute, Budapest, Hungary
H. Kimura Osaka University, Suita, Japan
L. Kreindlin Technion-Israel Institute of Technology, Technion City, Haifa, Israel
G. Kuesslin Universität Kassel, Kassel, F.R.G.
V. Kufner Czechoslovak Academy of Sciences, Prague, Czechoslovakia
H. Kwakernaak Univ., Philadelphia, PA, U.S.A.
R. Lofgren Univ. of California, Irvine, U.S.A.
M. Y. Marmel Univ. of Newcastle, NSW, Australia
J. Mitrani Mihailo Pupin Institute, Belgrade, Yugoslavia
R. V. Patel Concordia Univ., Montreal, Canada
M. G. Rodd Univ. College of Swansea, Swansea, U.K.
J. B. Sheridan Massachusetts Institute of Technology, Cambridge, MA, U.S.A.
T. Soderstrom Uppsala Univ., Uppsala, Sweden
Y. Sunahara Kyoto Institute of Technology, Kyoto, Japan
D. Tabak George Mason University, Fairfax, VA, U.S.A.
H. Unbehauen Ruhr Univ., Bochum, Bochum, F.R.G.
V. Vukobratovic Institute of Problems in Control, Moscow, U.S.S.R.
I. Valavanis Massachusetts Institute of Technology, Cambridge, MA, U.S.A.
G. Varghese Massachusetts Institute of Technology, Cambridge, MA, U.S.A.
B. Wahlberg Linköping Univ., Sweden
B. Wittenmark Lund Institute of Technology, Lund, Sweden
M. I. Younis American Univ. in Cairo, Egypt

Publishing, Subscription and Advertising Offices

Subscription enquiries from customers in North America should be sent to Pergamon Press Inc., Maxwell House, Fairview Park, Elmsford, NY 10523, U.S.A. and for the remainder of the world to Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K.

Subscription Rates (including postage and insurance)

Annual institutional subscription rate (1991-92): DM 875.00; 2 year institutional rate (1991-92): DM 1682.50. Personal subscription rate for IFAC Affiliates (1991): DM 100.00. Prices are subject to change without notice. Subscription rates for Japan include despatch by air and prices are available on application. Six issues per annum. Copyright © 1990 International Federation of Automatic Control (IFAC). For information about becoming an IFAC Affiliate, contact the IFAC Secretariat.

It is a condition of publication that manuscripts submitted to this journal have not been published and will not be simultaneously submitted or published elsewhere. By submitting a manuscript, the author agrees that the copyright for their article is transferred to IFAC if and when the article is accepted for publication. However, assignment of copyright is not required from authors who work for organizations which do not permit such assignment. The copyright covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microfilm or any other reproductions of similar nature and translations. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, electrostatic, magnetic tape, mechanical photocopying, recording or otherwise, without permission in writing from the copyright holder.

Photocopying information for users in the U.S.A. The Item Fee Code for this publication indicates that authorization to photocopy items for internal or personal use is granted by the copyright holder for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service provided the stated fee for copying beyond that permitted by Section 107 or 108 of the United States Copyright Law is paid. The appropriate remittance of \$3.00 per copy per article is paid directly to the Copyright Clearance Center Inc., 27 Congress Street, Salem, MA 01970.

Permission for other use. The copyright owner's consent does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific written permission must be obtained from the publisher for such copying.

The Item Fee Code for this publication is 0005-1098/91 \$3.00 + 0.00

Microform Subscriptions and Back Issues

Back issues of all previously published volumes, in both hard copy and on microform, are available direct from Pergamon Press offices.

♾️ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences: Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Continued from outside front cover

M. Dahleh and M. A. Dahleh	201	On Slowly Time-varying Systems
		BOOK REVIEWS
D. W. Clarke	207	Adaptive Control by K. J. Åström and B. Wittenmark
J. F. Bohme	209	Introduction to Signals and Systems by E. Kamen
J. L. Willems	210	State Variable Methods in Automatic Control by K. Furuta and A. Sano
R. Johansson	210	Binäre Steuerungstechnik—Eine Einführung by K. H. Fasol
R. Kuřhavý	211	Statistical Analysis and Control of Dynamic Systems by H. Akaike and T. Nakagawa
J. Michálek	213	Engineering Applications of Stochastic Processes: Theory, Problems and Solutions by A. Zayezdny, D. Tabak and D. Wulich
	215	Biographical Notes on Contributors to this Issue
	224	Addendum to Lists of Reviewers for Automatica

INDEXED/ABSTRACTED IN: *Appl. Mech. Rev.*, *Curr. Cont. ASCA*, *Aqua Abstr.*, *Cam. Sci. Abstr.*, *Curr. Cont. CompuMath.*, *Curr. Cont. Eng. Tech. & Applied Sci.*, *Comput. Cont.*, *Eng. Ind.*, *INSPEC Data.*, *Info. Sci. Abstr.*, *Math. R.*, *Oper. Res. Manage. Sci.*, *PASCAL-CNRS Data.*, *Curr. Cont. Sci. Cit. Ind.*, *Curr. Cont. SCISEARCH Data.*, *SSSA/CISA/ECA/ISMEC*, *Zentralblatt für Mathematik*

ISSN 0005-1098
ATCAA9 27(1) 1-224 (1991)



PUBLISHED BY

Pergamon Press

OXFORD · NEW YORK

BEIJING · FRANKFURT · SÃO PAULO · SEOUL · SYDNEY · TOKYO

PRINTED IN GREAT BRITAIN BY BPCC WHEATONS LTD, EXETER

Information for Contributors to Automatica

FORMAT

- 1 Submission of papers:** Manuscripts in English with (a) An English title (10 words desired max. length) (b) Author's name and author's affiliation including present address as a footnote. (c) **Abstract:** Must be on first page less than 200 words for Papers, 100 words for Brief Papers and 75 words for Technical Communiques
- 2 Forms of contributions:** Contributions should include (a) 5 copies of Papers and Brief Papers to be submitted to the appropriate Editor, and 1 copy to the Editor-in-Chief, 3 copies of Technical Communiques and Correspondence to be submitted to the Editor of Technical Communiques and Correspondence on good quality white paper or high quality xerographic copies thereof, the manuscript to be typed with double spacing throughout with adequate margins (4 cm). (b) *Good quality* original figures, biographies and author's photographs are *not to be sent* until requested. This material is not normally returned to authors even if rejected, but it will be treated confidentially. Manuscripts should be prepared according to the order: **Title page** (including suggested running title if the title is over 5 words, and name and address to whom correspondence and reprints should be sent). Papers must include the following features: **Abstract**, **Introduction** (to explain background work, nature and purpose of paper), **Body** (to contain primary message with clear lines of thought and of mathematical expressions, formulae to be well spaced out identifying all Greek letters and unusual symbols by name in the margin), **Conclusion section** (to indicate significant contribution with its limitations, advantages and possible applications), **Acknowledgements** (when appropriate), **References**, **Appendices** (with short titles as needed to explain development details[†]), figure captions and table captions which must be typed on separate pages, tables and figures. **Maximum length**[†] Papers: 25 double-spaced typed pages with no more than 16 figures. Brief Papers: 10 double-spaced typed pages and with no more than 8 figures. Technical Communiques: 6 double-spaced typed pages. **These maximum lengths are calculated assuming a word density of 300 words/page. If a typewriter or word processor producing a higher density of words/page is used then the maximum number of pages must be reduced accordingly.**
- 3 Style of contribution:** All tables, figures and equations to be numbered with Arabic numerals. In the text the words 'equation' and 'figure' to be typed as 'equation' and 'Fig.'. Avoid hyphenation at the end of a line. Symbols denoting vectors to be indicated for bold type by a wavy underline as follows: \underline{x} , \underline{y} . Weights and measurements should be expressed in metric units. All non-standard abbreviations or symbols to be defined when first mentioned.
- 4 Tables:** Tables for publication in the Journal may be reproduced direct from the author's typescript if suitable and will be treated in the same way as line diagrams. They should be submitted along with the figures and descriptive captions and footnotes should accompany those for the figures typed on a separate sheet. **A very clear top copy** of the tables should be submitted and large or long tables should be typed on continuing sheets. Each table to be indicated on the upper right hand corner. *No facilities exist at the Editorial Office for retyping.* In case of difficulty please consult the photoreprographic unit of your institution.
- 5 Figures:** All photographs, schemes and diagrams are to be referred to as figures and should be numbered consecutively and not included in the typescript. On the back of the figure the author should write his name, the figure number and an indication of the orientation of the figure. Line diagrams should be of a quality suitable for direct reproduction (*photocopies, blueprints and dyes are not acceptable*) and *no larger* than 22 × 28 cm. They should be drawn boldly in black ink on tracing film or white cartridge paper. The lettering should be between 1.5 and 3 mm in height *after* the diagram has been reduced in size for printing (it is desirable that figures are drawn so they will reduce to the single column width of 7.5 cm). Typewritten lettering does not reproduce satisfactorily. Photographs should be restricted to the minimum necessary and submitted as glossy prints. Descriptive captions to the figures should be typed on a separate sheet together with the figure number. *Adherence to these instructions will facilitate a swift publication time.*
- 6 References:** In the text the surname of the author(s) followed by the year of publication of the reference is given, e.g. 'It has been shown (Smith, 1964) that...' or 'Smith (1964) has shown...'. In case there are several publications by the same author(s) in the same year, use notations '1964a', '1964b', etc. Up to two authors can be mentioned in text references, three or more authors should be shortened to the first name with *et al.* References should comply with the abbreviated title of the journal as given in *World List*.
References should be listed at the end of the manuscript, arranged alphabetically by first author and for each author chronologically. *All references listed must be cited in the text at an appropriate point.* The form of listed references is as follows:
Abell, B. C. (1945). The examination of cell nuclei. *Biochem. J.* **35**, 123-126.
Abell, B. C., R. G. Tagg and M. Rush (1954). Enzyme catalyzed cellular transmission. In A. F. Round (Ed.) *Advances in Enzymology*, Vol. 2, pp. 125-247. 3rd ed. Academic Press, New York.
References, including those pending publication in well known journals or pertaining to private communications, not readily available to referees and readers will not be acceptable if the understanding of any part of the submitted paper is dependent upon them.
- 7 Biographies:** A short biography (to be written as paragraphs in narrative form) and passport-style photograph should be provided* with the figures. Authors of previous papers are reminded that biographies already published are not normally reproducible.
- 8 Packing:** Authors are advised to use a padded envelope (or 'Jiffy bag') when posting multiple copies of their paper. This should be well secured with strong adhesive tape. Single copies of papers should be posted only in strong manilla envelopes sealed so that the paper cannot move to and fro inside. Airmail should be used whenever appropriate. Packages sent to certain countries require a small green international customs label: the entries on this label should read 'printed material' or 'manuscript' or 'scientific paper' and 'no commercial value' or 'NCV'.
- 9 Proofs:** Unless otherwise specified, page proofs will be sent to the first named author for correction. Alterations should be avoided, apart from typographical errors made by the printer. Additional alterations will delay publication and if excessive may be charged to the author. *Proofs should be returned with 48 hours of receipt.*
- 10 Reprints:** Reprints can be purchased at a reasonable cost if ordered when proofs are returned on a reprint order form which accompanies page proofs. It would be appreciated if authors notify the publisher of any change of address occurring whilst their article is in the process of publication.
- 11 Responsibility for the contents of the paper rests upon the authors, and not upon IFAC, the editors, or the publisher.** Therefore, it is important that the technical content of the manuscript be carefully considered by the author.
It is also important that the author agrees not to publish the same manuscript in another Journal without obtaining the consent of the editor of this Journal.
- 12** The original manuscript and diagrams will be discarded 1 month after publication unless the publisher is requested to return the original material to the author.

*Does not apply to Technical Communiques

[†]On the basis of about 300 words per typed page and 2.5 figures equivalent to a typed page



automatica

The Journal of IFAC the International Federation of Automatic Control

Volume 27 Number 2

March 1991

CONTENTS

PAPERS

- | | | |
|---|-----|--|
| D. M. Perdu and A. H. Levis | 225 | A Petri Net Model for Evaluation of Expert Systems in Organizations |
| J. Prock | 239 | A New Technique for Fault Detection Using Petri Nets |
| I. Kanellakopoulos, P. V. Kokotovic and R. Marino | 247 | An Extended Direct Scheme for Robust Adaptive Nonlinear Control |
| B. M. Chen, A. Saberi and P. Sannuti | 257 | A New Stable Compensator Design for Exact and Approximate Loop Transfer Recovery |
| R. H. Middleton | 281 | Trade-offs in Linear Control System Design |
| H. Ozbay and A. Tannenbaum | 293 | On the Structure of Suboptimal H^∞ Controllers in the Sensitivity Minimization Problem for Distributed Stable Plants |
| M. A. Rotea and P. P. Khargonekar | 307 | H^2-optimal Control with an H^∞-constraint: The State Feedback Case |
| J. S. McDonald and J. B. Pearson | 317 | H^1-optimal Control of Multivariable Systems with Output Norm Constraints |
| C. Commault, J. M. Dion and J. A. Torres | 331 | Minimal Structure in the Block Decoupling Problem with Stability |
| M. Kárný | 339 | Estimation of the Control Period for Self-tuners |
| V. L. Syrmos and F. L. Lewis | 349 | A Geometric Approach to Proportional-plus-derivative Feedback Using Quotient and Partitioned Subspaces |
| J. Kornylo | 371 | BRIEF PAPERS
Kalman Smoothing via Auxiliary Outputs |
| R. V. Pate and P. Misra | 375 | Numerical Computation of Decentralized Fixed Modes |
| M. K. Sundareshan and R. M. Elbanna | 383 | Qualitative Analysis and Decentralized Controller Synthesis for a Class of Large-scale Systems with Symmetrically Interconnected Subsystems |
| W. Feng | 389 | On Practical Stability of Linear Multivariable Feedback Systems with Time-delays |
| K. J. Hunt and M. Šebek | 395 | Implied Polynomial Matrix Equations in Multivariable Stochastic Optimal Control |
| F. Giri, J. M. Dion, L. Dugard and M. M'Saad | 399 | Parameter Estimation Aspects in Adaptive Control |
| M. Milanese and A. Vicino | 403 | Estimation Theory for Nonlinear Models and Set Membership Uncertainty |

Continued on outside back cover
Pergamon Press

OXFORD · NEW YORK

automatica

The IFAC Journal

By an agreement between IFAC and Pergamon Press plc (the official IFAC publisher), AUTOMATICA is the official Journal of IFAC, the International Federation of Automatic Control

IFAC Council

President B. D. O. Anderson
President elect S. J. Kahne
Vice president L. Ljung
Vice president Y. Z. Lu
Immediate Past President B. Tamm
Treasurer M. Mansour
Ordinary Members
J. Ackermann
A. van Cauwenberghe
E. J. Davison
A. Ichikawa
V. Kucera
P. M. Larsen
A. Tili
J. D. N. van Wyk

Technical Board

Chairman L. Ljung
Vice chairmen
E. J. Davison
L. Keviczky
P. M. Larsen
M. G. Rodd
P. Urosen

Executive Board

Chairman Y. Z. Lu
Immediate Past President B. Tamm
Treasurer M. Mansour
Secretary G. Hencsey
Chairman Policy Committee P. M. Larsen
Chairman Publication Committee G. Guardabassi
President elect S. J. Kahne

Editorial Board

Chairman G. S. Axelby
Vice chairmen
H. A. Spang III
H. Kwakernaak
Chairman Publication Committee G. Guardabassi
Chairman of IFAC Publications
Managing Board M. Thoma
Members
K. J. Åström
P. M. Larsen
W. S. Levine
P. C. Parks
M. Rodd (Technical Board Liaison)
A. P. Sage
P. T. Shepherd
R. E. Strange

IFAC Secretariat

G. Hencsey
B. Aumann
E. Rulas
Schlossplatz 12
2301 Lauenburg
Austria

Parameter Estimation and Adaptive Control

Patrick C. Parks
Mathematics Group, School of Defence Management
Royal Military College of Science
Shrivenham, Swindon, SN6 6EA, U.K.

Large-scale Systems, Management and Decision Sciences

Andrew P. Sage
George Mason University
4400 University Drive
Fairfax, VA 22030, U.S.A.

Survey Papers

Karl J. Åström
Division of Automatic Control
Lund Institute of Technology
S-221 00 Lund 7, Sweden

Technical Communiques and Correspondence, Rapid Publications

William S. Levine
Dept of Electrical Engineering
University of Maryland, MD 20742, U.S.A.

State Estimation, Optimal Control and Systems

Hubert Kwakernaak
Faculty of Applied Mathematics
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands

Computer Control and Applications

H. Austin Spang III
G.E. CR&D
Box 8, KWD 220, Schenectady
NY 12301, U.S.A.

Book Reviews

Peter Martin Larsen
Electrical Power Engineering Dept
Bldg 325, Technical University of Denmark
2800 Lyngby, Denmark

Authors should send five copies of manuscripts for publication to appropriate Editor and one copy to the Editor-in-Chief with copy of letter to Editor.

Associate Editors

- Y. Arkun, Georgia Institute of Technology, Atlanta, GA, U.S.A.
K. E. Årzén, Lund Institute of Technology, Lund, Sweden
A. R. B. Chao, Univ. of Twente, The Netherlands
T. Beyer, Univ. of Illinois, IL, U.S.A.
H. Bittmann, Australian National Univ., Canberra, Australia
H. K. Bui, Univ. of Liège, Belgium
R. Canales, Inst. Tecnológico de Chihuahua, Chihuahua, Mexico
A. van Cauwenberghe, Rijksuniversiteit Gent, Zwijnaarde, Belgium
J. H. Chow, G.E. ESDO, Schenectady, NY, U.S.A.
H. W. Clarke, Univ. of Oxford, Oxford, U.K.
R. Corbett, Univ. of Groningen, The Netherlands
P. Dorato, Univ. of New Mexico, Albuquerque, NM, U.S.A.
B. Friedland, New Jersey Institute of Technology, Newark, NJ, U.S.A.
P. J. Gawthrop, The University, Glasgow, U.K.
G. Guardabassi, Politecnico di Milano, Milano, Italy
Y. Y. Haimm, Univ. of Virginia, VA, U.S.A.
C. C. Hang, Nat'l Univ. of Singapore, Singapore
M. Ikeda, Aichi University, Japan
H. Isenmann, Technische Hochschule Darmstadt, F.R.G.
A. Isodon, Univ. di Roma, La Sapienza, Rome, Italy
M. Johnson, Univ. of Strathclyde, Glasgow, U.K.
S. J. Kahne, The MITRE Corp., McLean, VA, U.S.A.
L. Kaszkurewicz, Rio de Janeiro Federal Univ., Rio de Janeiro, Brazil
L. Keviczky, Computers and Automation Institute, Budapest, Hungary
H. Kimura, Osaka University, Suita, Japan
E. Kresel, Technion-Israel Institute of Technology, Technion City, Haifa, Israel
V. Kulakov, Czechoslovak Academy of Sciences, Prague, Czechoslovakia
H. Kwakernaak, Univ. of Twente, Enschede, The Netherlands
H. Lugo, Univ. of Chile, Chile
I. M. Y. Mareels, Australian National Univ., Canberra, Australia
J. Medanic, Mihailo Pupin Institute, Belgrade, Yugoslavia
H. V. Patel, Concordia Univ., Montreal, Canada
M. G. Rodd, Univ. College of Swansea, Swansea, U.K.
T. B. Shenton, Massachusetts Institute of Technology, Cambridge, MA, U.S.A.
T. Soderstrom, Uppsala Univ., Uppsala, Sweden
Y. Sunahara, Kyoto Institute of Technology, Kyoto, Japan
D. Tabak, George Mason University, Fairfax, VA, U.S.A.
H. Unbehauen, Ruhr Univ. Bochum, Bochum, F.R.G.
V. I. Utkin, Institute of Problems in Control, Moscow, U.S.S.R.
T. Vlasov, Massachusetts Institute of Technology, Cambridge, MA, U.S.A.
G. Verghese, Massachusetts Institute of Technology, Cambridge, MA, U.S.A.
B. Wahlberg, Linköping Univ., Sweden
H. Wittenmark, Lund Institute of Technology, Lund, Sweden
M. I. Younis, American Univ. in Cairo, Egypt

This staff of Associate Editors is being expanded to include representatives from various countries throughout the world.

Publishing, Subscription and Advertising Offices

Subscription requests from customers in North America should be sent to: Pergamon Press Inc., Maxwell House, Fairview Park, Elmsford, NY 10523, U.S.A. and for the remainder of the world to: Pergamon Press plc, Headington Hill Hall, Oxford, OX3 0BW, U.K.

Subscription Rates (including postage and insurance)

Annual institutional subscription rate (1991): DM 875.00; 2 year institutional rate (1991-92): DM 1662.50. Personal subscription rate for IFAC Affiliates (1991): DM 100.00. Prices are subject to change without notice. Subscription rates for Japan include postage by air and prices are available on application. Six issues per annum. Copyright © 1991, International Federation of Automatic Control (IFAC). For information about becoming an IFAC Affiliate, contact the IFAC Secretariat.

It is a condition of publication that manuscripts submitted to this journal have not been published and will not be simultaneously submitted or published elsewhere. By submitting a manuscript, the authors agree that the copyright for their article is transferred to IFAC if and when the article is accepted for publication. However, assignment of copyright is not required from authors who work for organizations which do not permit such assignment. The copyright covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microform or any other reproductions of similar nature and translations. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, electrostatic, magnetic tape, mechanical photocopying, recording or otherwise, without permission in writing from the copyright holder.

Whilst every effort is made by the publishers and editorial board to see that no inaccurate or misleading data, opinion or statement appears in this journal, they wish to make it clear that the data and opinions appearing in the articles and advertisements herein are the sole responsibility of the contributor or advertiser concerned. Accordingly, the publishers, the editorial board and editors and their respective employees, officers and agents accept no responsibility or liability whatsoever for the consequences of any such inaccurate or misleading data, opinion or statement.

Photocopying information for users in the U.S.A. The Item Fee Code for this publication indicates that authorization to photocopy items for internal or personal use is granted by the copyright holder for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service provided the stated fee for copying beyond that permitted by Section 107 or 108 of the United States Copyright Law is paid. The appropriate remittance of \$3.00 per copy per article is paid directly to the Copyright Clearance Center Inc., 27 Congress Street, Salem, MA 01970.

Permission for other use. The copyright owner's consent does not extend to copying for general distribution for promotion, for creating new works, or for resale. Specific written permission must be obtained from the publisher for such copying.

The Item Fee Code for this publication is: 0005-1098/91 \$3.00 + 0.00

Microform Subscriptions and Back Issues

Back issues of all previously published volumes in both hard copy and on microform are available direct from Pergamon Press offices.

♾️ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-19.

Automatica, Vol. 27, No. 2

Continued from outside front cover

B. M. Mohan and K. B. Datta	409	Linear Time-invariant Distributed Parameter System Identification via Orthogonal Functions
L. Baratchart, M. Cardelli and M. Oliveri	413	Identification and Rational L^2 Approximation: A Gradient Algorithm
M. S. Chiu and Y. Arkun	419	TECHNICAL COMMUNIQUE A New Result on Relative Gain Array, Niederlinski Index and Decentralized Stability Condition: 2 x 2 Plant Cases
R. Ortega	423	Passivity Properties for Stabilization of Cascaded Nonlinear Systems
S. Kannila and T. Westerlund	425	An Elementary Derivation of the Maximum Likelihood Estimator of the Covariance Matrix, and an Illustrative Determinant Inequality
L. Motus	427	BOOK REVIEWS <i>Real-time Computer Control: An Introduction</i> by S. Bennett
C. Schmid	428	<i>Computer Control of Machines and Processes</i> by J. G. Bollinger and N. A. Duffei
H. Rake	429	<i>Industrial Control Electronics: Applications and Design</i> by J. M. Jacob
H. Ramon	430	<i>Power Hydraulics</i> by M. J. Pinches and J. G. Ashby
J. F. Barrett	431	<i>Random Signals and Systems</i> by R. E. Mortensen
	435	<i>Biographical Notes on Contributors to this Issue</i>

INDEXED/ABSTRACTED IN *Appl Mech Rev*, *Curr Cont ASCA*, *Aqua Abstr*, *Cam Sci Abstr*, *Curr Cont CompuMath*, *Curr Cont/Eng Tech & Applied Sci*, *Comput. Cont.*, *Eng. Ind.*, *INSPEC Data*, *Info Sci Abstr*, *Math R*, *Oper Res Manage Sci*, *PASCAL-CNRS Data*, *Curr Cont Sci. Cit. Ind.*, *Curr Cont SCISEARCH Data*, *SSSA/CISA/ECA/ISMEC*, *Zentralblatt fur Mathematik*

ISSN 0005-1098
ATCAA9 27(2) 225-440 (1991)



PUBLISHED BY

Pergamon Press

OXFORD NEW YORK

FRANKFURT SEOUL SYDNEY TOKYO

PRINTED IN GREAT BRITAIN BY BPCC WHEATONS LTD, EXETER

Information for Contributors to Automatica

FORMAT

- 1 Submission of papers.** Manuscripts in English with (a) An English title (10 words desired max. length) (b) Author's name and author's affiliation including present address as a footnote (c) **Abstract** - Must be on first page less than 200 words for Papers, 100 words for Brief Papers and 75 words for Technical Communiques
- 2 Forms of contributions** Contributions should include (a) 5 copies of Papers and Brief Papers to be submitted to the appropriate Editor, and 1 copy to the Editor-in-Chief, 3 copies of Technical Communiques and Correspondence to be submitted to the Editor of Technical Communiques and Correspondence on good quality white paper or high quality xerographic copies thereof, the manuscript to be typed with double spacing throughout with adequate margins (4 cm) (b) *Good quality* original figures, biographies and author's photographs are *not to be sent* until requested. This material is not normally returned to authors even if rejected, but it will be treated confidentially. Manuscripts should be prepared according to the order: *Title page* (including suggested running title if the title is over 5 words, and name and address to whom correspondence and reprints should be sent) *Papers* must include the following features: *Abstract*, *Introduction* (to explain background work, nature and purpose of paper), *Body* (to contain primary message with clear lines of thought and of mathematical expressions - formulae to be well spaced out identifying all Greek letters and unusual symbols by name in the margin), *Conclusion section* (to indicate significant contribution with its limitations, advantages and possible applications), *Acknowledgements* (when appropriate), *References*, *Appendices* (with short titles as needed to explain development details*), figure captions, and table captions which must be typed on separate pages, tables and figures. *Maximum length* † *Papers* - 25 double-spaced typed pages with no more than 16 figures. *Brief Papers* - 10 double-spaced typed pages and with no more than 8 figures. *Technical Communiques* - 6 double-spaced typed pages. ***These maximum lengths are calculated assuming a word density of 300 words/page. If a typewriter or word processor producing a higher density of words/page is used then the maximum number of pages must be reduced accordingly***
- 3 Style of contribution.** All tables, figures and equations to be numbered with Arabic numerals. In the text the words 'equation' and 'figure' to be typed as 'equation' and 'Fig.'. Avoid hyphenation at the end of a line. Symbols denoting vectors to be indicated for bold type by a wavy underline as follows: $\underline{\underline{x}}$, $\underline{\underline{y}}$. Weights and measurements should be expressed in metric units. All non-standard abbreviations or symbols to be defined when first mentioned.
- 4 Tables.** Tables for publication in the Journal may be reproduced direct from the author's typescript if suitable and will be treated in the same way as line diagrams. They should be submitted along with the figures and descriptive captions and footnotes should accompany those for the figures *typed on a separate sheet*. **A very clear top copy** of the tables should be submitted and large or long tables should be typed on continuing sheets. Each table to be indicated on the upper right hand corner. ***No facilities exist at the Editorial Office for retyping.*** In case of difficulty please consult the photoreprographic unit of your institution.
- 5 Figures.** All photographs, schemes and diagrams are to be referred to as figures and should be numbered consecutively and not included in the typescript. On the back of the figure the author should write his name, the figure number and an indication of the orientation of the figure. Line diagrams should be of a quality suitable for direct reproduction (*photocopies, blueprints and dylines are not acceptable*) and *no larger* than 22 x 28 cm. They should be drawn boldly in black ink on tracing film or white cartridge paper. The lettering should be between 1.5 and 3 mm in height *after* the diagram has been reduced in size for printing (it is desirable that figures are drawn so they will reduce to the single column width of 7.5 cm). Typewritten lettering does not reproduce satisfactorily. Photographs should be restricted to the minimum necessary and submitted as glossy prints. Descriptive captions to the figures should be typed on a separate sheet together with the figure number. ***Adherence to these instructions will facilitate a swift publication time***
- 6 References.** In the text the surname of the author(s) followed by the year of publication of the reference is given, e.g. 'It has been shown (Smith, 1964) that ...' or 'Smith (1964) has shown ...'. In case there are several publications by the same author(s) in the same year, use notations 1964a, 1964b, etc. Up to two authors can be mentioned in text references, three or more authors should be shortened to the first name with *et al.* References should comply with the abbreviated title of the journal as given in *World List*.
References should be listed at the end of the manuscript, arranged alphabetically by first author and for each author chronologically. ***All references listed must be cited in the text at an appropriate point.*** The form of listed references is as follows:
Abell, B. C. (1945) The examination of cell nuclei. *Biochem. J.* **35**, 123-126.
Abell, B. C., R. G. Tagg and M. Rush (1954) Enzyme catalyzed cellular transmission. In A. F. Round (Ed.) *Advances in Enzymology*, Vol. 2, pp. 125-247. 3rd ed. Academic Press, New York.
References including those pending publication in well known journals or pertaining to private communications, not readily available to referees and readers will not be acceptable if the understanding of any part of the submitted paper is dependent upon them.
- 7 Biographies.** A short biography (to be written as paragraphs in narrative form) and passport-style photograph should be provided* with the figures. Authors of previous papers are reminded that biographies already published are not normally reproducible.
- 8 Packing.** Authors are advised to use a padded envelope (or 'Jiffy bag') when posting multiple copies of their paper. This should be well secured with strong adhesive tape. Single copies of papers should be posted only in strong manilla envelopes sealed so that the paper cannot move to and fro inside. Airmail should be used whenever appropriate. Packages sent to certain countries require a small green international customs label: the entries on this label should read 'printed material' or 'manuscript' or 'scientific paper' and 'no commercial value' or 'NCV'.
- 9 Proofs.** Unless otherwise specified, page proofs will be sent to the first-named author for correction. Alterations should be avoided, apart from typographical errors made by the printer. Additional alterations will delay publication and, if excessive, may be charged to the author. ***Proofs should be returned with 48 hours of receipt***
- 10 Reprints.** Reprints can be purchased at a reasonable cost if ordered when proofs are returned on a reprint order form which accompanies page proofs. It would be appreciated if authors notify the publisher of any change of address occurring whilst their article is in the process of publication.
- 11 Responsibility for the contents of the paper rests upon the authors, and not upon IFAC, the editors, or the publisher.** Therefore it is important that the technical content of the manuscript be carefully considered by the author.
It is also important that the author agrees not to publish the same manuscript in another Journal without obtaining the consent of the editor of this Journal.
- 12** The original manuscript and diagrams will be discarded 1 month after publication unless the publisher is requested to return the original material to the author.

*Does not apply to Technical Communiques

†On the basis of about 300 words per typed page and 2.5 figures equivalent to a typed page

automatica

The Journal of IFAC the International Federation of Automatic Control

Volume 27 Number 3

May 1991

CONTENTS

	PAPERS
P. M. Mills, P. L. Lee and P. McIntosh	441 A Practical Study of Adaptive Control of an Alumina Calciner
R. M. Stephan, V. Hahn, J. Dastych and H. Unbehauen	449 Adaptive and Robust Cascade Schemes for Thyristor Driven DC-motor Speed Control
J. Levine and P. Rouchon	463 Quality Control of Binary Distillation Columns via Nonlinear Aggregated Models
P. Tsotras and H. J. Kelley	481 Drag-law Effects in the Goddard Problem
J. Bentsman, K. S. Hong and J. Fakfakh	491 Vibrational Control of Nonlinear Time Lag Systems. Vibrational Stabilization and Transient Behavior
G. Fruchter	501 Generalized Zero Sets Location and Absolute Robust Stabilization of Continuous Nonlinear Control Systems
	BRIEF PAPERS
K. Akimoto, N. Sannomiya, Y. Nishikawa and T. Tsuda	513 An Optimal Gas Supply for a Power Plant Using a Mixed Integer Programming Model
J. H. Lee and M. Morari	519 Robust Measurement Selection
C. C. Hang and D. Chin	529 Reduced Order Process Modelling in Self-tuning Control
A. De Luca, L. Lanari and G. Oriolo	535 A Sensitivity Approach to Optimal Spline Robot Trajectories
W. A. Berger, R. J. Perry and H. H. Sun	541 An Algorithm for the Assignment of System Zeros
W. L. Chen and J. S. Gibson	545 A Lyapunov Robustness Bound for Linear Systems with Periodic Uncertainties
J. C. Hennet and J.-P. Beziat	549 A Class of Invariant Regulators for the Discrete-time Linear Constrained Regulation Problem
A. Casavola, M. J. Grimble, E. Mosca and P. Nistri	555 Continuous-time LQ Regulator Design by Polynomial Equations
J. S. Shamma and M. Athans	559 Guaranteed Properties of Gain Scheduled Control for Linear Parameter-varying Plants
D. Yaniv	565 Arbitrarily Small Sensitivity in Multiple-input-output Uncertain Feedback Systems
J. R. Partington	569 Approximation of Delay Systems by Fourier-Laguerre Series

Continued on outside back cover



Pergamon Press

OXFORD · NEW YORK

FRANKFURT · SEQUOIA · SYDNEY · TOKYO

The IFAC Journal

By an agreement between IFAC and Pergamon Press plc (the official IFAC publisher), AUTOMATICA is the official Journal of IFAC, the International Federation of Automatic Control

IFAC Council

President B. D. O. Anderson
President-elect S. J. Kahne
Vice-president L. Ljung
Vice-president Y. Z. Lu
Immediate Past President B. Tamm
Treasurer M. Mansour
Ordinary Members
J. Ackermann
A. van Cauwenbergh
E. J. Davison
A. Ichikawa
V. Kucera
P. M. Larsen
A. Titli
J. D. N. van Wyk

Technical Board

Chairman L. Ljung
Vice-chairmen
E. J. Davison
I. Kavcivsky
P. M. Larsen
M. G. Rodd
P. Uraosen

Executive Board

Chairman Y. Z. Lu
Immediate Past President B. Tamm
Treasurer M. Mansour
Secretary G. Hencsey
Chairman Policy Committee P. M. Larsen
Chairman Publication Committee G. Guardabassi
President-elect S. J. Kahne

Editorial Board

Chairman G. S. Axelby
Vice-chairmen
H. A. Spang III
H. Kwakernaak
Chairman Publication Committee G. Guardabassi
Chairman of IFAC Publications
Managing Board M. Thoma
Members
K. J. Åström
P. M. Larsen
W. S. Levine
P. C. Parks
M. Rodd (Technical Board Liaison)
A. P. Sage
P. T. Shepherd
R. E. Strange

IFAC Secretariat

G. Hencsey
A. Aumann
R. Ruda
Schlossplatz 12
301 Laxenburg
Austria

Editor-in-Chief

George S. Axelby
Automatica
211 Coronet Drive
North Linthicum
MD 21090, U.S.A.
Tel: (301) 789-0284
Telex 9102406716
Fax: (301) 636-8625

Parameter Estimation and Adaptive Control

Patrick C. Parks
Mathematics Group, School of Defence Management
Royal Military College of Science
Beverham, Swindon, SN6 8LA, U.K.

State Estimation, Optimal Control and Systems

Huibert Kwakernaak
Faculty of Applied Mathematics
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands

Large-scale Systems, Management and Decision Sciences

Andrew P. Sage
George Mason University
400 University Drive
 Fairfax, VA 22030, U.S.A.

Computer Control and Applications

H. Austin Spang III
G.E. CR&D
Box 8, KWD 220, Schenectady
NY 12301, U.S.A.

Survey Papers

Karl J. Åström
Division of Automatic Control
Lund Institute of Technology
S-221 00 Lund 7, Sweden

Technical Communiques and Correspondence, Rapid Publications

William S. Levine
Dept of Electrical Engineering
University of Maryland, MD 20742, U.S.A.

Book Reviews

Peer Martin Larsen
Electrical Power Engineering Dept
Bldg 325 Technical University of Denmark
2800 Lyngby, Denmark

Authors should send five copies of manuscripts for publication to appropriate Editor and one copy to the Editor-in-Chief with copy of letter to Editor.

Associate Editors

Y. Arkun Georgia Institute of Technology, Atlanta, GA, U.S.A.
K. E. Årzen Lund Institute of Technology, Lund, Sweden
A. Baychi Univ. of Twente, The Netherlands
T. Başar Univ. of Illinois, U.S.A.
R. Bitmead Australian National Univ., Canberra, Australia
R. K. Boel Univ. of Gent, Belgium
H. Canales Ruiz Gobernador de Chihuahua, El Collado del Mirador, Cuernavaca, Mexico
A. van Cauwenbergh Rijksuniversiteit Gent, Zwijnaarde, Belgium
J. H. Chow G.E. EUSD, Schenectady, NY, U.S.A.
D. W. Clarke Univ. of Oxford, Oxford, U.K.
R. Curtain Univ. of Groningen, The Netherlands
P. Dorato Univ. of New Mexico, Albuquerque, NM, U.S.A.
B. Friedland New Jersey Institute of Technology, Newark, NJ, U.S.A.
P. J. Gawthrop The University, Glasgow, U.K.
G. Guardabassi Politecnico di Milano, Milano, Italy
Y. Y. Haimm Univ. of Virginia, VA, U.S.A.
C. C. Hang Nat'l Univ. of Singapore, Singapore
M. Ikeda Kobe University, Japan
R. Isenmann Technische Hochschule Darmstadt, F.R.G.
A. Isidori Univ. di Roma, La Sapienza, Rome, Italy
M. Johnson Univ. of Strathclyde, Glasgow, U.K.
S. J. Kahne The MITRE Corp., McLean, VA, U.S.A.
E. Kaszkurewicz Rio de Janeiro Federal Univ., Rio de Janeiro, Brazil

This staff of Associate Editors is being expanded to include representatives from various countries throughout the world.

Publishing, Subscription and Advertising Offices

Subscription enquiries from customers in North America should be sent to Pergamon Press Inc., Maxwell House, Fairview Park, Elmsford, NY 10523, U.S.A. and for the remainder of the world to Pergamon Press plc, Headington Hill Hall, Oxford, OX3 0BW, U.K.

Subscription Rates (including postage and insurance)

Annual institutional subscription rate (1991) DM 875.00; 2-year institutional rate (1991-92) DM 1662.50. Personal subscription rate for IFAC Affiliates (1991) DM 100.00. Prices are subject to change without notice. Subscription rates for Japan include despatch by air and prices are available on application. Six issues per annum. Copyright © 1991 International Federation of Automatic Control (IFAC). For information about becoming an IFAC Affiliate, contact the IFAC Secretariat.

It is a condition of publication that manuscripts submitted to this journal have not been published and will not be simultaneously submitted or published elsewhere. By submitting a manuscript, the authors agree that the copyright for their article is transferred to IFAC, and when the article is accepted for publication, however, assignment of copyright is not required from authors who work for organizations which do not permit such assignment. The copyright covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microform or any other reproductions of similar nature and translations. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, electrostatic, magnetic tape, mechanical, photocopying, recording or otherwise, without permission in writing from the copyright holder.

Whilst every effort is made by the publishers and editorial board to see that no inaccurate or misleading data, opinion or statement appears in this journal, they wish to make it clear that the data and opinions appearing in the articles and advertisements herein are the sole responsibility of the contributor or advertiser concerned. Accordingly, the publishers, the editorial board and editors and their respective employees, officers and agents accept no responsibility or liability whatsoever for the consequences of any such inaccurate or misleading data, opinion or statement.

Photocopying information for users in the U.S.A.: The item fee code for this publication indicates that authorization to photocopy items for internal or personal use is granted by the copyright holder or libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service provided the stated fee for copying, beyond that permitted by Section 107 or 108 of the United States Copyright Law, is paid. The appropriate remittance of \$3.00 per copy per article is paid directly to the Copyright Clearance Center Inc., 27 Congress Street, Salem, MA 01970.

Permission for other use: The copyright owner's consent does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific written permission must be obtained from the publisher for such copying.

The item fee code for this publication is: 0006-1098/91 \$3.00 + 0.00

Microform Subscriptions and Back Issues

Back issues of all previously published volumes, in both hard copy and on microform, are available direct from Pergamon Press offices.

♾️ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Continued from outside front cover

B. Rustem	573	The Diagonalizability of Quadratic Functions and the Arbitrariness of Shadow Prices
Y. Y. Haines and D. Li	579	A Hierarchical-multiobjective Framework for Risk Management
R. Jiroušek	585	BOOK REVIEWS <i>Expert Systems: Principles and Programming</i> by J. C. Giarratano and G. Riley
E. Gottzein	586	<i>Large Space Structures: Dynamics and Control</i> by S. N. Atluri and A. K. Amos
T. Vámos	588	<i>Applied Control of Manipulation Robots: Analysis, Synthesis and Exercises</i> by M. Vukobratovic and D. Stokic <i>Applied Dynamics of Manipulation Robots: Modelling, Analysis and Examples</i> by M. Vukobratovic
I. M. Y. Mareels	590	<i>Dynamic Models and Discrete Event Simulation</i> by W. Delaney and E. Vaccari
	591	<i>Biographical Notes on Contributors to this Issue</i>

INDEXED/ABSTRACTED IN: *Appl. Mech. Rev.*, *Curr. Cont. ASCA*, *Aqua Abstr.*, *Cam. Sci. Abstr.*, *Curr. Cont. CompuMath.*, *Curr. Cont. Eng. Tech. & Applied Sci.*, *Comput. Cont. Eng. Ind.*, *INSPEC Data*, *Info. Sci. Abstr.*, *Math. R. Oper. Res. Manage. Sci.*, *PASCAL*, *CNRS Data*, *Curr. Cont. Sci. Cit. Ind.*, *Curr. Cont. SCISEARCH Data*, *SSSA*, *CISA*, *ECA*, *ISMEC*, *Zentralblatt für Mathematik*

ISSN 0005-1098
ATCAA9 27(3) 441-598 (1991)



PUBLISHED BY

Pergamon Press

OXFORD NEW YORK

FRANKFURT SEOUL SYDNEY TOKYO

PRINTED IN GREAT BRITAIN BY BPCC WHEATONS LTD. EXETER

Information for Contributors to Automatica

FORMAT

- 1 **Submission of papers:** Manuscripts in English with (a) An English title (10 words desired max length) (b) Author's name and author's affiliation including present address as a footnote (c) **Abstract**—Must be on first page less than 200 words for Papers, 100 words for Brief Papers and 75 words for Technical Communiques
- 2 **Forms of contributions:** Contributions should include (a) 5 copies of Papers and Brief Papers to be submitted to the appropriate Editor, and 1 copy to the Editor-in-Chief, 3 copies of Technical Communiques and Correspondence to be submitted to the Editor of Technical Communiques and Correspondence on good quality white paper or high quality xerographic copies thereof, the manuscript to be typed with double spacing throughout with adequate margins (4 cm) (b) *Good quality* original figures, biographies and author's photographs are *not to be sent* until requested. This material is not normally returned to authors even if rejected, but it will be treated confidentially. Manuscripts should be prepared according to the order—*Title page* (including suggested running title if the title is over 5 words, and name and address to whom correspondence and reprints should be sent), Papers must include the following features: *Abstract*, *Introduction* (to explain background work, nature and purpose of paper), *Body* (to contain primary message with clear lines of thought and of mathematical expressions—formulae to be well spaced out identifying all Greek letters and unusual symbols by name in the margin), *Conclusion section* (to indicate significant contribution with its limitations, advantages and possible applications), *Acknowledgements* (when appropriate), *References*, *Appendices* (with short titles as needed to explain development details*), figure captions, and table captions which must be typed on separate pages, tables and figures. *Maximum length*† Papers—25 double-spaced typed pages with no more than 16 figures. Brief Papers—10 double-spaced typed pages and with no more than 8 figures. Technical Communiques—6 double-spaced typed pages. ***These maximum lengths are calculated assuming a word density of 300 words/page. If a typewriter or word processor producing a higher density of words/page is used then the maximum number of pages must be reduced accordingly***
- 3 **Style of contribution:** All tables, figures and equations to be numbered with Arabic numerals. In the text the words 'equation' and 'figure' to be typed as 'equation' and 'Fig'. Avoid hyphenation at the end of a line. Symbols denoting vectors to be indicated for bold type by a wavy underline as follows \underline{x} , \underline{y} . Weights and measurements should be expressed in metric units. All non-standard abbreviations or symbols to be defined when first mentioned
- 4 **Tables:** Tables for publication in the Journal may be reproduced direct from the author's typescript if suitable and will be treated in the same way as line diagrams. They should be submitted along with the figures and descriptive captions and footnotes should accompany those for the figures *typed on a separate sheet*. A *very clear top copy* of the tables should be submitted and large or long tables should be typed on continuing sheets. Each table to be indicated on the upper right-hand corner. *No facilities exist at the Editorial Office for retyping*. In case of difficulty please consult the photoreprographic unit of your institution
- 5 **Figures:** All photographs, schemes and diagrams are to be referred to as figures and should be numbered consecutively and not included in the typescript. On the back of the figure the author should write his name, the figure number and an indication of the orientation of the figure. Line diagrams should be of a quality suitable for direct reproduction (*photocopies, blueprints and dylines are not acceptable*) and *no larger* than 22 × 28 cm. They should be drawn boldly in black ink on tracing film or white cartridge paper. The lettering should be between 1.5 and 3 mm in height *after* the diagram has been reduced in size for printing (it is desirable that figures are drawn so they will reduce to the single column width of 7.5 cm). Typewritten lettering does not reproduce satisfactorily. Photographs should be restricted to the minimum necessary and submitted as glossy prints. Descriptive captions to the figures should be typed on a separate sheet together with the figure number. *Adherence to these instructions will facilitate a swift publication time*
- 6 **References:** In the text the surname of the author(s) followed by the year of publication of the reference is given, e.g. 'It has been shown (Smith, 1964) that...' or 'Smith (1964) has shown...'. In case there are several publications by the same author(s) in the same year, use notations '1964a', '1964b', etc. Up to two authors can be mentioned in text references, three or more authors should be shortened to the first name with *et al*. References should comply with the abbreviated title of the journal as given in *World List*.
References should be listed at the end of the manuscript, arranged alphabetically by first author and for each author chronologically. *All references listed must be cited in the text at an appropriate point*. The form of listed references is as follows:
Abell, B. C. (1945) The examination of cell nuclei. *Biochem. J.* **35**, 123–126.
Abell, B. C., R. G. Tagg and M. Rush (1954) Enzyme-catalyzed cellular transmission. In A. F. Round (Ed.), *Advances in Enzymology*, Vol. 2, pp. 125–247. 3rd ed. Academic Press, New York.
References, including those pending publication in well-known journals or pertaining to private communications, not readily available to referees and readers will not be acceptable if the understanding of any part of the submitted paper is dependent upon them.
- 7 **Biographies:** A short biography (to be written as paragraphs in narrative form) and passport-style photograph should be provided* *with the figures*. Authors of previous papers are reminded that biographies already published are not normally reproducible
- 8 **Packing:** Authors are advised to use a padded envelope (or 'Jiffy bag') when posting multiple copies of their paper. This should be well-secured with strong adhesive tape. Single copies of papers should be posted only in strong manila envelopes sealed so that the paper cannot move to and fro inside. Airmail should be used whenever appropriate. Packages sent to certain countries require a small green international customs label: the entries on this label should read 'printed material' or 'manuscript' or 'scientific paper' and 'no commercial value' or 'NCV'.
- 9 **Proofs:** Unless otherwise specified, page proofs will be sent to the first-named author for correction. Alterations should be avoided, apart from typographical errors made by the printer. Additional alterations will delay publication and, if excessive, may be charged to the author. *Proofs should be returned with 48 hours of receipt*.
- 10 **Reprints:** Reprints can be purchased at a reasonable cost if ordered when proofs are returned on a reprint order form which accompanies page proofs. It would be appreciated if authors notify the publisher of any change of address occurring whilst their article is in the process of publication.
- 11 **Responsibility for the contents of the paper rests upon the authors, and not upon IFAC, the editors, or the publisher.** Therefore, it is important that the technical content of the manuscript be carefully considered by the author.
It is also important that the author agrees not to publish the same manuscript in another Journal without obtaining the consent of the editor of this Journal.
- 12 The original manuscript and diagrams will be discarded 1 month after publication unless the publisher is requested to return the original material to the author.

*Does not apply to Technical Communiques

†On the basis of about 300 words per typed page and 2.5 figures equivalent to a typed page.

Call for Papers: *Automatica* Special Issue on Robust Control

IN THE 1980s robustness became one of the most popular subjects in control theory research. Robustness is a fundamental issue in feedback control, and indeed one of the main reasons for using feedback. Because of its central importance, *Automatica* plans to publish a Special Issue on Robust Control, covering the following themes:

Robustness analysis. Structured and unstructured perturbations, stability and performance robustness.

Robustness design. Methods and algorithms for the design of robust control systems.

Case studies. State-of-the-art design studies of practical robust feedback systems.

Although adaptive control is much concerned with robustness, it is a subject by itself that is excluded from the scope of this Special Issue. Optimization papers, such as on H_∞ -optimization, will be selected on the basis of their relevance for robustness design.

Well-known authors are being invited to prepare tutorial papers surveying special areas, but there will be ample space for contributed papers, both in the form of regular and of brief papers.

The Special Issue will be prepared by a team

consisting of

Guest editors: J. Ackermann,
German Aerospace
Establishment
P. Dorato,
University of New Mexico
B. A. Francis,
University of Toronto
Automatica R. E. Curtain,
editorial staff University of Groningen
H. Kimura,
Osaka University
H. Kwakernaak,
University of Twente

Authors are invited to submit papers to
H. Kwakernaak, Editor of *Automatica*
University of Twente
P. O. Box 217, 7500 AE Enschede
The Netherlands
(Tel.) intl + 31-53-903457
(Fax) intl + 31-53-340733
E-mail: twhuib@utwente.nl

according to the following time schedule:

Submission deadline: 1 September 1991
Final selection of papers: 1 March 1992
Special issue: January 1993

Authors are encouraged to submit papers early or to announce their intention of submitting a paper well before the deadline.

Survey Paper

Some Aspects of Parallel and Distributed Iterative Algorithms—A Survey*†

DIMITRI P. BERTSEKAS‡§ and JOHN N. TSITSIKLIS‡

Iterative methods suitable for use in parallel and distributed computing systems are surveyed. Both synchronous and asynchronous implementations are discussed. A number of theoretical issues regarding the validity of asynchronous algorithms are addressed.

Key Words—Computational methods; distributed data processing; iterative methods; parallel processing; asynchronous algorithms; parallel algorithms; distributed algorithms

Abstract—We consider iterative algorithms of the form $x_{k+1} = f(x_k)$, executed by a parallel or distributed computing system. We first consider synchronous executions of such iterations and study their communication requirements, as well as issues related to processor synchronization. We also discuss the parallelization of iterations of the Gauss-Seidel type. We then consider asynchronous implementations whereby each processor iterates on a different component of x_k at its own pace, using the most recently received (but possibly outdated) information on the remaining components of x_k . While certain algorithms may fail to converge when implemented asynchronously, a large number of positive convergence results is available. We classify asynchronous algorithms into three main categories, depending on the amount of asynchronism they can tolerate, and survey the corresponding convergence results. We also discuss issues related to their termination.

1. INTRODUCTION

PARALLEL AND DISTRIBUTED computing systems have received broad attention motivated by several different types of applications. Roughly speaking, *parallel* computing systems consist of several tightly coupled processors that are located within a small distance of each other. Their main purpose is to execute jointly a computational task and they have been designed with such a purpose in mind; communication between processors is fast and reliable. *Distributed* computing systems are somewhat

different in a number of respects. Processors are loosely coupled with little, if any, central coordination and control, and interprocessor communication is more problematic. Communication delays can be unpredictable, and the communication links themselves can be unreliable. Finally, while the architecture of a parallel system is usually chosen with a particular set of computational tasks in mind, the structure of distributed systems is often dictated by exogenous considerations. Nevertheless, there are several algorithmic issues that arise in both parallel and distributed systems and that can be addressed jointly. To avoid repetition, we will mostly employ in the sequel the term “distributed”, but it should be kept in mind that most of the discussion applies to parallel systems as well.

There are at least two contexts where distributed computation has played a significant role. The first is the context of information acquisition, information extraction, and control, within spatially distributed systems. An example is a sensor network in which a set of geographically distributed sensors obtain information on the state of the environment and process it cooperatively. Another example is provided by data communication networks in which certain functions of the network (such as correct and timely routing of messages) have to be controlled in a distributed manner, through the cooperation of the computers residing at the nodes of the network. Other applications are possible in the quasistatic decentralized control of large scale systems whereby certain parameters (e.g. operating points for each subsystem) are to be optimized locally, while taking into account interactions with neighboring subsystems. The second important context for parallel or distributed computation is the solution of very large computational problems in which no single processor has sufficient computational power to tackle the problem on its own.

The ideas of this paper are relevant to both contexts, but our presentation will emphasize large scale numerical computation issues and iterative methods in particular. Accordingly, we shall consider

* Received 10 November 1988; revised 12 February 1990; received in final form 11 March 1990. The original version of this paper was presented at the IFAC/IMACS Symposium on Distributed Intelligence Systems which was held in Varna, Bulgaria during June 1988. The published Proceedings of this IFAC Meeting may be ordered from Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Editor K. J. Åström.

† Research supported by the NSF under Grants ECS-8519058 and ECS-8552419, with matching funds from Bellcore, Du Pont and IBM, and by the ARO under Grant DAAL03-86-K-0171.

‡ Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

§ Author to whom all correspondence should be addressed.

algorithms of the form $x := f(x)$ where $x = (x_1, \dots, x_n)$ is a vector in \mathcal{R}^n and $f: \mathcal{R}^n \rightarrow \mathcal{R}^n$ is an iteration mapping defining the algorithm. In many interesting applications, it is natural to consider distributed executions of this iteration whereby the i th processor updates x_i according to the formula

$$x_i := f_i(x_1, \dots, x_n), \quad (1.1)$$

while receiving information from other processors on the current values of the remaining components.

Our discussion of distributed implementations of iteration (1.1) focuses on mechanisms for interprocessor communication and synchronization. We also consider asynchronous implementations and present a survey of the convergence issues that arise in the face of asynchronism. These issues are discussed in more detail in Bertsekas and Tsitsiklis (1989b) where proofs of most of the results quoted here can be found.

Iteration (1.1) can be executed *synchronously* whereby processors perform an iteration, communicate their results to the other processors, and then proceed to the next iteration. In Section 2, we introduce two alternative synchronous iterations, namely Jacobi type and Gauss-Seidel type iterations, and discuss briefly their parallelization. In Section 3, we indicate that synchronous parallel execution is feasible even if the underlying computing system is inherently asynchronous (i.e. no processor has access to a global clock) provided that certain synchronization mechanisms are in place. We review and compare three representative synchronization methods. We also discuss some basic communication problems that arise naturally in parallel iterations, assuming that processors communicate using a point-to-point communication network. Then, in Section 4, we provide a more detailed analysis of the required time per parallel iteration. In Section 5, we indicate that the synchronous execution of iteration (1.1) can have certain drawbacks, thus motivating *asynchronous* implementations whereby each processor computes at its own pace while receiving (possibly outdated) information on the values of the components updated by the other processors. An asynchronous implementation of iteration (1.1) is not mathematically equivalent to its synchronous counterpart and an otherwise convergent algorithm may become divergent. It will be seen that asynchronous iterative algorithms can display several and different convergence behaviors, ranging from divergence to guaranteed convergence in the face of the worst possible amount of asynchronism and communication delays. We classify the possible behaviors in three broad classes, the corresponding convergence results are surveyed in Sections 6, 7 and 8, respectively. In Section 9, we address some difficulties that arise if we wish to terminate an asynchronous distributed algorithm in finite time. Finally, Section 10 contains our conclusions and a brief discussion of future research directions.

2. JACOBI AND GAUSS-SEIDEL ITERATIONS

Let X_1, \dots, X_p be subsets of the Euclidean spaces $\mathcal{R}^{n_1}, \dots, \mathcal{R}^{n_p}$, respectively. Let $n = n_1 + \dots + n_p$,

and let $X \subset \mathcal{R}^n$ be the Cartesian product $X = \prod_{i=1}^p X_i$.

Accordingly, any $x \in \mathcal{R}^n$ is decomposed in the form $x = (x_1, \dots, x_p)$, with each x_i belonging to \mathcal{R}^{n_i} . For $i = 1, \dots, p$, let $f_i: X \rightarrow X_i$ be a given function and let $f: X \rightarrow X$ be the function defined by $f(x) = (f_1(x), \dots, f_p(x))$ for every $x \in X$. We want to solve the fixed point problem $x = f(x)$. To this end we will consider the iteration

$$x := f(x).$$

We will also consider the more general iteration

$$x_i := \begin{cases} f_i(x) & \text{if } i \in I \\ x_i & \text{otherwise,} \end{cases} \quad (2.1)$$

where I is a subset of the component index set $\{1, \dots, p\}$, which may change from one iteration to the next.

We are interested in the distributed implementation of such iterations. While some of the discussion applies to shared memory systems, we will focus in this and the next two sections on a message-passing system with p processors, each having its own local memory and communicating with the other processors over a communication network. We assume that the i th processor has the responsibility of updating the i th component x_i according to the rule $x_i := f_i(x)$. It is implicitly assumed here that the i th processor knows the form of the function f_i . In the special case where $f(x) = Ax + b$, where A is an $n \times n$ matrix and $b \in \mathcal{R}^n$, this amounts to assuming that the i th processor knows the *rows* of the matrix A corresponding to the components assigned to it. Other implementations of the linear iteration $x := Ax + b$ are also possible. For example, each processor could be given certain *columns* of A . We do not pursue this issue further and refer the reader to McBryan and Van der Velde (1987) and Fox *et al.* (1988) for discussions of alternative matrix storage schemes.

For implementation of the iteration, it is seen that if the function f_i depends on x_j (with $i \neq j$), then processor j must be informed by processor i on the current value of x_i . To capture such data dependencies, we form a directed graph $G = (N, A)$, called the *dependency graph* of the algorithm, with nodes $N = \{1, \dots, p\}$ and with arcs $A = \{(i, j) \mid i \neq j \text{ and } f_i \text{ depends on } x_j\}$. We assume that for every arc (i, j) in the dependency graph there is a communication capability by means of which processor i can relay information to processor j . We also assume that messages are received correctly within a finite but otherwise arbitrary amount of time. Such communication may be possible through a direct communication link joining processors i and j or it could consist of a multi-hop path in a communication network. The discussion that follows applies to both cases.

An iteration in which all of the components of x are simultaneously updated [$I = \{1, \dots, p\}$ in (2.1)], is sometimes called a *Jacobi* type iteration. In an alternative form, the components of x are updated one at a time, and the most recently computed values of the other components are used. The resulting iteration is often called an iteration of the

Gauss-Seidel type and is described mathematically by

$$\begin{aligned} x_i(t+1) &= f_i(x_1(t+1), \dots, x_{i-1}(t+1), \\ &\quad x_i(t), \dots, x_p(t)), \\ &\quad i = 1, \dots, p \end{aligned} \quad (2.2)$$

In a serial computing environment, Gauss-Seidel iterations are often preferable. As an example, consider the linear case where $f(x) = Ax + b$, and A has non-negative elements and spectral radius less than one. Then, the classical Stein Rosenberg theorem [see e.g. Bertsekas and Tsitsiklis (1989b, p. 152)] states that both the Gauss-Seidel and the Jacobi iterations converge at a geometric rate to the unique fixed point of f , however, in a serial setting where one Jacobi iteration takes as much as one Gauss-Seidel iteration, the rate of convergence of the Gauss-Seidel iteration is always faster. Surprisingly, in a parallel setting this conclusion is reversed, as we now describe in a somewhat more general context.

Consider the sequence $\{x^j(t)\}$ generated by the Jacobi iteration

$$x^j(t+1) = f(x^j(t)), \quad t = 0, \quad (2.3)$$

and the sequence $\{x^{j*}(t)\}$ generated by the Gauss-Seidel iteration (2.2), started from the same initial condition $x(0) = x^j(0) = x^{j*}(0)$. The following result is proved in Tsitsiklis (1989) generalizing an earlier result of Smart and White (1988).

Proposition 1. Suppose that $f: \mathcal{R}^n \rightarrow \mathcal{R}^n$ has a unique fixed point x^* , and is monotone, that is, it satisfies $f(x) \leq f(y)$ if $x \leq y$. Then, if $f(x(0)) \leq x(0)$, we have

$$x^* \leq x^j(pt) \leq x^{j*}(t), \quad t = 0, 1, \dots$$

and if $x(0) \geq f(x(0))$, we have

$$x^{j*}(t) \leq x^j(pt) \leq x^*, \quad t = 0, 1, \dots$$

Proposition 1 establishes the faster parallel convergence of the Jacobi iteration, for certain initial conditions, assuming that a Gauss-Seidel iteration takes as much parallel time as p Jacobi iterations. It has also been shown in Smart and White (1988) that if in addition to the assumptions of Proposition 2.1, f is linear (and thus satisfies the assumptions of the Stein-Rosenberg theorem), the rate of convergence of the Jacobi iteration is faster than the rate of convergence of the Gauss-Seidel iteration. An extension of this result that applies to asynchronous Jacobi and Gauss-Seidel iterations is also given in Bertsekas and Tsitsiklis (1989a).

The preceding comparison of Jacobi and Gauss-Seidel iterations assumes that a Jacobi iteration is executed in one time step, and that the Gauss-Seidel iteration cannot be parallelized (so that a full update of all the components x_1, \dots, x_p requires p time steps). This is the case when the number of available processors is p and the dependency graph describing the structure of the iteration is complete (every component depends on every other component), so that no two components can be updated in parallel. A Gauss-Seidel iteration can still converge faster, however, if it can be parallelized to the point where it requires the same number of time steps as the

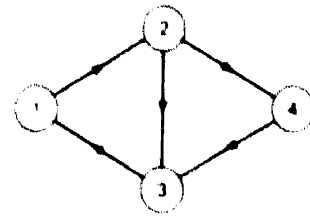


FIG. 1. A dependency graph

corresponding Jacobi iteration; this can happen if the number of available processors is less than p and the dependency graph is sufficiently sparse, as we now illustrate.

Consider the dependency graph of Fig. 1. A corresponding Gauss-Seidel iteration is described by

$$\begin{aligned} x_1(t+1) &= f_1(x_1(t), x_3(t)) \\ x_2(t+1) &= f_2(x_1(t+1), x_3(t)) \\ x_3(t+1) &= f_3(x_1(t+1), x_3(t), x_4(t)) \\ x_4(t+1) &= f_4(x_2(t+1), x_3(t)) \end{aligned}$$

and its structure is shown in Fig. 2. We notice here that $x_1(t+1)$ and $x_4(t+1)$ can be computed in parallel. In particular, a sweep, that is, an update of all four components, can be performed in only three stages. On the other hand, a different ordering of the components leads to an iteration of the form

$$\begin{aligned} x_1(t+1) &= f_1(x_1(t), x_3(t)) \\ x_3(t+1) &= f_3(x_1(t), x_3(t), x_4(t)) \\ x_4(t+1) &= f_4(x_1(t), x_4(t)) \\ x_2(t+1) &= f_2(x_1(t+1), x_3(t)) \end{aligned}$$

which is illustrated in Fig. 3. We notice here that $x_1(t+1)$, $x_3(t+1)$, and $x_4(t+1)$ can be computed in parallel, and a sweep requires only two stages.

The above example motivates the problem of choosing an ordering of the components for which a sweep requires the least number of stages. The solution of this problem, given in Bertsekas and Tsitsiklis (1989b, p. 23) is as follows:

Proposition 2. The following are equivalent:

- There exists an ordering of the variables such that a sweep of the corresponding Gauss-Seidel algorithm can be performed in K parallel steps.
- We can assign colors to the nodes of the dependency graph so that at most K different colors

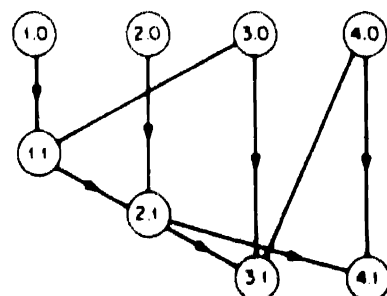


FIG. 2. The data dependencies in a Gauss-Seidel iteration.

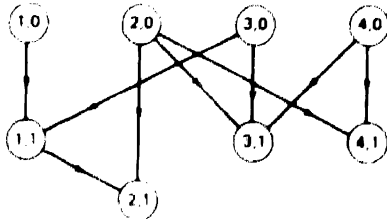


FIG. 3. The data dependencies in a Gauss-Seidel iteration for a different updating order.

are used and so that each subgraph obtained by restricting to the set of nodes with the same color has no directed cycles.

A well known special case of the above proposition arises when the dependency graph G is symmetric; that is, the presence of an arc $(i, j) \in A$ also implies the presence of the arc (j, i) . In this case there is no need to distinguish between directed and undirected cycles, and the coloring problem of Proposition 2 reduces to coloring the nodes of the dependency graph so that no two neighboring nodes have the same color.

Unfortunately, the coloring problem of Proposition 2 is intractable (NP-hard). On the other hand, in several practical situations the dependency graph G has a very simple structure and the coloring problem can be solved by inspection. Furthermore, it can be shown that if the dependency graph is a tree or a two-dimensional grid, only two colors suffice, so a Gauss-Seidel sweep can be done in two steps, with roughly half the components of x being updated in parallel at each step. In this case, while with n processors the Jacobi method is as fast or faster than Gauss-Seidel, the reverse is true when using $n/2$ processors (or more generally, any number of processors with which a Gauss-Seidel step can be completed in the same time as the Jacobi iteration).

Even with unstructured dependency graphs, reasonably good colorings can be found using simple heuristics; see Zenios and Lasken (1988) and Zenios and Mulvey (1988), for examples. Let us also point out that the parallelization of Gauss-Seidel methods by means of coloring is very common in the context of the numerical solution of partial differential equations; see, for example, Ortega and Voigt (1985) and the references therein.

A related approach for parallelizing Gauss-Seidel iterations, which is fairly easy to implement, is discussed in Barbosa (1986) and Barbosa and Gafni (1987). In this approach, a new sweep is allowed to start before the previous one has been completed and for this reason, one obtains, in general, somewhat greater parallelism than that obtained by the coloring approach.

We finally note that the order in which the variables are updated in a Gauss-Seidel sweep may have a significant effect on the convergence rate of the iteration. Thus, completing a Gauss-Seidel sweep in a minimum number of steps is not the only consideration in selecting the grouping of variables to be

updated in parallel; the corresponding rate of convergence must also be taken into account.

3. SYNCHRONIZATION AND COMMUNICATION ISSUES

We say that an execution of iteration (2.1) is *synchronous* if it can be described mathematically by the formula

$$x_i(t+1) = \begin{cases} f_i(x_1(t), \dots, x_p(t)) & \text{if } i \in T' \\ x_i(t) & \text{otherwise.} \end{cases} \quad (3.1)$$

Here, t is an integer-valued variable used to index different iterations, not necessarily representing real time, and T' is an infinite subset of the index set $\{0, 1, \dots\}$. Thus, T' is the set of time indices at which x_i is updated. With different choices of T' one obtains different algorithms, including Jacobi and Gauss-Seidel type of methods. We will later contrast synchronous iterations with asynchronous iterations, where instead of the current component values $x_j(t)$, earlier values $x_j(t-d)$ are used in (3.1), with d being a possibly positive and unpredictable "communication delay" that depends on i, j and t .

3.1. Synchronization methods

Synchronous execution is certainly possible if the processors have access to a global clock, and if messages can be reliably transmitted from one processor to another between two consecutive "ticks" of the clock. Barring the existence of a global clock, synchronous execution can be still accomplished by using synchronization protocols called *synchronizers*. We refer the reader to Awerbuch (1985) for a comparative complexity analysis of a class of synchronizers and we continue with a brief discussion of three representative synchronization methods. These methods will be described for the case of Jacobi type iterations, but they can be easily adapted for the case of Gauss-Seidel iterations as well.

(a) *Global synchronization.* Here the processors proceed to the $(t+1)$ st iteration, also referred to as phase, only after every processor i has completed the t th iteration and has received the value of $x_j(t)$ from every j such that $(j, i) \in A$. Global synchronization can be implemented by a variety of techniques, a simple one being the following: the processors are arranged as a spanning tree, with a particular processor chosen to be the root of the tree. Once processor i has computed $x_i(t)$, has received the value of $x_j(t)$ for every j such that $(j, i) \in A$, and has received a phase termination message from all its "children" in the tree, it sends a phase termination message to its "father" in the tree. Phase termination messages thus propagate towards the root. Once the root has received a phase termination message from all of its children, it knows that the current phase has been completed and sends a phase initiation message to its children, which is propagated along the spanning tree. Once a processor receives such a message it can proceed to the next phase. (See Fig. 4 for an illustration.)

(b) *Local synchronization.* Global synchronization can be seen to be rather wasteful in terms of the time

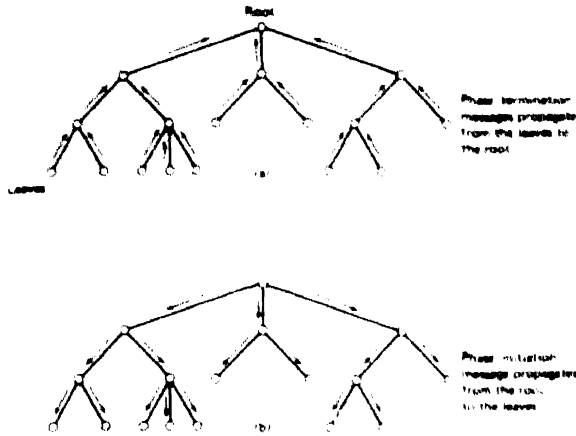


FIG. 4. Illustration of the global synchronization method

required per iteration. An alternative is to allow the i th processor to proceed with the $(t+1)$ st iteration as soon as it has received all the messages $x_j(t)$ it needs. Thus, processor i moves ahead on the basis of local information alone, obviating the need for propagating messages along a spanning tree.

It is easily seen that the iterative computation can only proceed faster when local synchronization is employed. Furthermore, this conclusion can also be reached even if a more efficient global synchronization method were possible whereby all processors start the $(t+1)$ st iteration immediately after all messages generated by the t th iteration have been delivered. (We refer to this hypothetical and practically unachievable situation as the ideal global synchronization.) Let us assume that the time required for on-computation and the communication delays are bounded above by a finite constant and are bounded below by a positive constant. Then it is easily shown that the time spent for a number K of iterations under ideal global synchronization is at most a constant multiple of the corresponding time when local synchronization is employed.

The advantage of local synchronization is better seen if communication delays do not obey any *a priori* bound. For example, let us assume that the communication delay of every message is an independent exponentially distributed random variable with mean one. Furthermore, suppose for simplicity, that each processor sends messages to exactly d other processors, where d is some constant (i.e. the outdegree of each node of the dependency graph is equal to d). With global synchronization, the real time spent for one iteration is roughly equal to the maximum of dp independent exponential random variables and its expectation is, therefore, of the order of $\log(dp)$. Thus, the expected time needed for K iterations is of the order of $K \log(pd)$. On the other hand, with local synchronization, it turns out that the expected time for K iterations is of the order of $\log p + K \log d$ [joint work with C. H. Papadimitriou; see Bertsekas and Tsitsiklis (1989b, p. 104)]. If K is large, then local synchronization is faster by a factor roughly equal to $\log(pd)/\log d$. Its advantage is more

pronounced if d is much smaller than p , as is the case in most practical applications. Some related analysis and experiments can be found in Dubois and Briggs (1982).

(c) *Synchronization via rollback.* This method, introduced by Jefferson (1985), has been primarily applied to the simulation of discrete-event systems. It can also be viewed as a general purpose synchronization method but it is likely to be inferior to the preceding two methods in applications involving solution of systems of equations. Consider a situation where the message $x_j(t)$ transmitted from some processor j to some other processor i is most likely to take a fixed default value known to i . In such a case, processor i may go ahead with the computation of $x_i(t+1)$ without waiting for the value of $x_j(t)$, by making the assumption that $x_j(t)$ will take the default value. In case that a message comes later which falsifies the assumption that $x_j(t)$ has the default value, then a *rollback* occurs: that is, the computation of $x_i(t+1)$ is invalidated and is performed once more, taking into account the correct value of $x_j(t)$. Furthermore, if a processor has sent messages based on computations which are later invalidated, it sends *antimessages* which cancel the earlier messages. A reception of such an antimessage by some other processor k could invalidate some of k 's computations and could trigger the transmission of further antimessages by k . This process has the potential of explosive generation of antimessages that could drain the available communication resources. On the other hand, it is hoped that the number of messages and antimessages would remain small in problems of practical interest, although insufficient analytical evidence is available at present. Some probabilistic analyses of the performance of this method can be found in Lavenberg *et al.* (1983) and Mitra and Mitram (1984).

3.2 Single and multinode broadcasting

Regardless of whether the implementation is synchronous or not, it is necessary to exchange some information between the processors after each iteration. The interprocessor communication time can be substantial when compared to the time devoted to computations, and it is important to carry out the message exchanges as efficiently as possible. There are a number of generic communication problems that arise frequently in iterative and other algorithms. We describe a few such tasks related to message broadcasting.

In the first communication task, we want to send the same message from a given processor to every other processor (we call this a *single node broadcast*). In a generalized version of this problem, we want to do a single node broadcast simultaneously from all nodes (we call this a *multinode broadcast*). A typical example where a multinode broadcast is needed arises in the iteration $x := f(x)$. If we assume that there is a separate processor assigned to component x_i , $i = 1, \dots, p$, and that the function f_j depends on all components x_i , $j = 1, \dots, p$, then, at the end of an iteration, there is a need for every processor to send

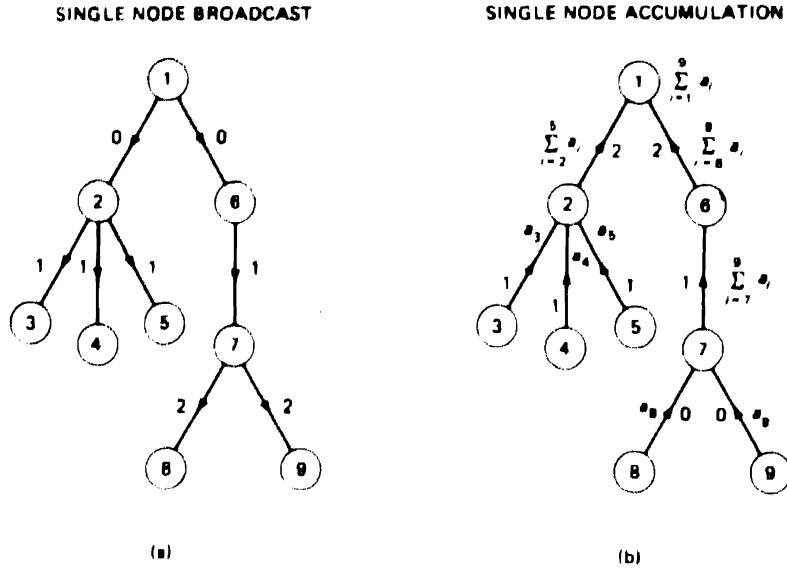


FIG. 5. (a) A single node broadcast uses a tree that is rooted at a given node (which is node 1 in the figure). The time next to each link is the time that transmission of the packet on the link begins. (b) A single node accumulation problem involving summation of n scalars a_1, \dots, a_n (one per processor) at the given node (which is node 1 in the figure). The time next to each link is the time at which transmission of the "combined" packet on the link begins, assuming that the time for scalar addition is negligible relative to the time required for packet transmission.

the value of its component to every other processor, which is a multinode broadcast.

Clearly, to solve the single node broadcast problem, it is sufficient to transmit the given node's message along a spanning tree rooted at the given node, that is, a spanning tree of the network together with a direction on each link of the tree such that there is a unique path from the given node (called the *root*) to every other node. With an optimal choice of such a spanning tree, a single node broadcast takes $\Theta(r)$ -time,[†] where r is the diameter of the network, as shown in Fig. 5(a). To solve the multinode broadcast problem, we need to specify one spanning tree per root node. The difficulty here is that some links may belong to several spanning trees; this complicates the timing analysis, because several messages can arrive simultaneously at a node, and require transmission on the same link with a queueing delay resulting.

There are two important communication problems that are dual to the single and multinode broadcasts, in the sense that the spanning tree(s) used to solve one problem can also be used to solve the dual in the same amount of communication time. In the first problem, called *single node accumulation*, we want to send to a given node a message from every other node; we assume, however, that messages can be "combined" for transmission on any communication link, with a "combined" transmission time equal to the transmission time of a single message. This problem arises, for example, when we want to form at a given node a sum consisting of one term for each

node, as in an inner product calculation [see Fig. 5(b)], we can view addition of scalars at a node as "combining" the corresponding messages into a single message. The second problem, which is dual to a multinode broadcast, is called *multinode accumulation*, and involves a separate single node accumulation at each node. It can be shown that a single node (or multinode) accumulation problem can be solved in the same time as a single node (respectively multinode) broadcast problem, by realizing that an accumulation algorithm can be viewed as a broadcast algorithm running in reverse time, as illustrated in Fig. 5. As shown in Fig. 4, global synchronization can be accomplished by a single node broadcast followed by a single node accumulation.

Algorithms for solving the broadcast problems just described, together with other related communication problems, have been developed for several popular architectures (Nassimi and Sahni, 1980; Saad and Shultz, 1987; McBryan and Van der Velde, 1987; Ozveren, 1987; Bertsekas *et al.*, 1989; Bertsekas and Tsitsiklis, 1989b; Johnsson and Ho, 1989). Table 1 gives the order of magnitude of the time needed to solve each of these problems using an optimal algorithm. The underlying assumption for the results of this table is that each message requires unit time for transmission on any link of the interconnection network, and that each processor can transmit and receive a message simultaneously on all of its incident links. Specific algorithms that attain these times are given in Bertsekas *et al.* (1989) and Bertsekas and Tsitsiklis (1989b, Section 1.3.4). In most cases these algorithms are optimal in that they solve the problem in the minimum possible number of time steps. Figure 6 illustrates a multinode broadcast algorithm for a ring

[†] The notation $h(y) = \Theta(g(y))$, where y is a positive integer, means that for some $c_1 > 0$, $c_2 > 0$, and $y_0 > 0$, we have $c_1|g(y)| \leq h(y) \leq c_2|g(y)|$ for all $y \geq y_0$.

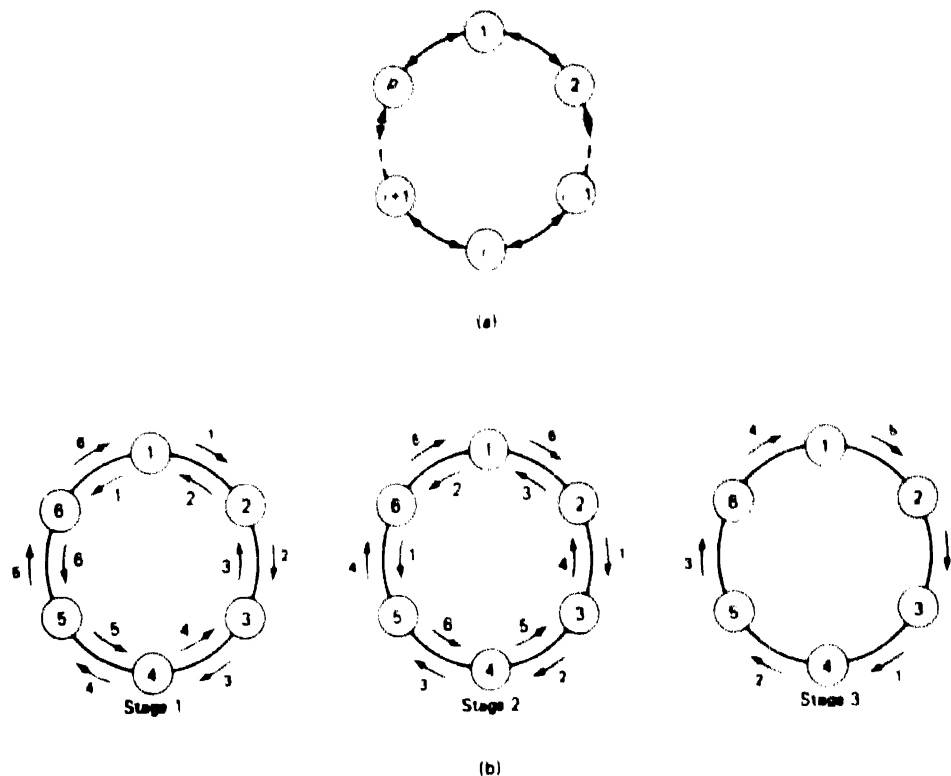


FIG. 6. (a) A ring of p nodes having as links the pairs $(i, i + 1)$ for $i = 1, 2, \dots, p - 1$, and $(p, 1)$. (b) A multinode broadcast on a ring with p nodes can be performed in $\lfloor (p - 1)/2 \rfloor$ stages as follows: at stage 1, each node sends its own packet to its clockwise and counterclockwise neighbors. At stage $2 \leq i \leq \lfloor (p - 1)/2 \rfloor$, each node sends to its clockwise neighbor the packet received from its counterclockwise neighbor at the previous stage; also, at stages $2 \leq i \leq \lfloor (p - 1)/2 \rfloor$, each node sends to its counterclockwise neighbor the packet received from its clockwise neighbor at the previous stage. The figure illustrates this process for $p = 6$.

with p processors, which attains the minimum number of steps.

Using the results of Table 1, it is also shown in Bertsekas and Tsitsiklis (1989b) that if a hypercube is used, then most of the basic operations of numerical linear algebra, i.e. inner product, matrix vector multiplication, matrix matrix multiplication, power of a matrix, etc., can be executed in parallel in the same order of time as when communication is instantaneous. In some cases this is also possible when the processors are connected with a less powerful interconnection network such as a square mesh. Thus, communication affects only the "multiplying constant" as opposed to the order of time needed to carry out these operations. Nonetheless, with a large number of processors, the effect of communication delays on linear algebra operations can be very substantial.

4. ITERATION COMPLEXITY

We now try to assess the potential benefit from parallelization of the iteration $x := f(x)$. In particular, we will estimate the order of growth of the required time per iteration, as the dimension n increases. Our analysis is geared towards large problems and the issue of speedup of iterative methods using a large number of processors. We will make the following

assumptions:

- (a) All components of x are updated at each iteration. (This corresponds to a Jacobi iteration. If a Gauss-Seidel iteration is used instead, the time per iteration cannot increase, since by updating only a subset of the components, the computation per iteration will be reduced and the communication problem will be simplified. Based on this, it can be seen that the order of required time will be unaffected if in place of a Jacobi iteration, we perform a Gauss-Seidel sweep with

TABLE 1. SOLUTION TIMES OF OPTIMAL ALGORITHMS FOR THE BROADCAST AND ACCUMULATION PROBLEMS USING A RING, A BINARY-BALANCED TREE, A d -DIMENSIONAL MESH (WITH THE SAME NUMBER OF PROCESSORS ALONG EACH DIMENSION) AND A HYPERCUBE WITH p PROCESSORS. THE TIMES GIVEN FOR THE RING ALSO HOLD FOR A LINEAR ARRAY.

Problem	Ring	Tree	Mesh	Hypercube
Single node broadcast (or single node accumulation)	$\Theta(p)$	$\Theta(\log p)$	$\Theta(p^{1/d})$	$\Theta(\log p)$
Multinode broadcast (or multinode accumulation)	$\Theta(p)$	$\Theta(p)$	$\Theta(p)$	$\Theta(p/\log p)$

a number of steps which is fixed and independent of the dimension n .)

- (b) There are n processors, each updating a single scalar component of x at each iteration. (One may wish to use fewer than n processors, say p , each updating an n/p -dimensional component of x , in order to economize on communication. We argue later, however, that under our assumptions, choosing $p < n$ cannot improve the order of time required per iteration, although it may reduce this time by a constant factor. In practice, of course, the number of available processors is often much less than n , and it is interesting to consider optimal utilization of a limited number of processors in the context of iterative methods. In this paper, however, we will not address this issue, preferring to concentrate on the potential and limitations of iterative computation using massively parallel machines with an abundant number of processors.)
- (c) Following the execution of their assigned portion of the iteration, the processors exchange the updated values of their components by means of a communication algorithm such as a multinode broadcast. The subsequent synchronization takes negligible time. (This can be justified by noting that local synchronization can be accomplished as part of the communication algorithm and thus requires no additional time. Furthermore, global synchronization can be done by means of a single node broadcast followed by a single node accumulation. Thus the time required for global synchronization grows with n no faster than a multinode broadcast time. Therefore, if the communication portion of the iteration is done by a multinode broadcast, the global synchronization time can be ignored when estimating the order of required time per iteration.)

We estimate the time per iteration as

$$T_{\text{COMP}} + T_{\text{MNB}},$$

where T_{COMP} is the time to compute the updated components $f_i(x)$, and T_{MNB} is the time to exchange the updated component values between the processors as necessary. If there is overlap of the computation and communication phases due to some form of pipelining, the time per iteration will be smaller than $T_{\text{COMP}} + T_{\text{MNB}}$ but its order of growth with n will not change. We consider several hypotheses for T_{COMP} and T_{MNB} , corresponding to different types of computation and communication hardware, and structures of the functions f_i . In particular, we consider the following cases, motivated primarily by the case where the system of equations $x = f(x)$ is linear:

Small T_{COMP} : ($= \Theta(1)$). One example for this case is when the iteration functions f_i are linear and correspond to a very sparse system (the maximum node degree of the dependency graph is $\Theta(1)$).

Another example is when the system solved is linear and dense, but each processor has vector processing capability allowing it to compute inner products in $\Theta(1)$ time.

Medium T_{COMP} : ($= \Theta(\log n)$). An example for this case is when the system solved is linear and dense, and each processor can compute an inner product in $\Theta(\log n)$ time. It can be shown that this is possible if each processor is itself a message-passing parallel processor with $\log n$ diameter.

Large T_{COMP} : ($= \Theta(n)$). An example for this case is when the system solved is linear and dense, and each processor computes inner products serially in $\Theta(n)$ time.

Also the following are considered for the communication time T_{MNB} :

Small T_{MNB} : ($= \Theta(1)$). An example for this case is when special very fast communication hardware is used, making the time for the multinode broadcast negligible relative to T_{COMP} or relative to the communication software overhead at the message sources. Another example is when the processors are connected by a network that matches the form of the dependency graph, so that all necessary communication involves directly connected nodes. For example when solving partial differential equations, the dependency graph is often a grid resulting from discretization of physical space. Then, with processors arranged in an appropriate grid, communication can be done very fast.

Medium T_{MNB} : ($= \Theta(n/\log n)$). An example for this case is when the multinode broadcast is performed using a hypercube network (cf. Table 1).

Large T_{MNB} : ($= \Theta(n)$). An example for this case is when the multinode broadcast is performed using a ring network or a linear array (cf. Table 1).

Table 2 gives the time per iteration $T_{\text{COMP}} + T_{\text{MNB}}$ for the different combinations of cases. In the worst case, the time per iteration is $\Theta(n)$, and this time is faster by a factor n than the time needed to execute serially the linear iteration $x := Ax + b$ when the matrix A is fully dense. In this case, the speedup is proportional to the number of processors n and the benefit from parallelization is very substantial. This thought, however, must be tempered by the realization that the parallel solution time still increases at least linearly with n , unless the number of iterations needed to solve the problem within practical accuracy decreases with n —an unlikely possibility.

TABLE 2. TIME PER ITERATION $x := f(x)$ UNDER A VARIETY OF ASSUMPTIONS FOR THE COMPUTATION TIME PER ITERATION T_{COMP} AND THE COMMUNICATION TIME PER ITERATION T_{MNB} . IN THE CELLS ABOVE THE DIAGONAL, THE COMPUTATION TIME IS THE BOTTLENECK, AND IN THE CELLS BELOW THE DIAGONAL, THE COMMUNICATION TIME IS THE BOTTLENECK

	$T_{\text{COMP}}: \Theta(1)$	$T_{\text{COMP}}: \Theta(\log n)$	$T_{\text{COMP}}: \Theta(n)$
$T_{\text{MNB}}: \Theta(1)$	$\Theta(1)$	$\Theta(\log n)$	$\Theta(n)$
$T_{\text{MNB}}: \Theta(n/\log n)$	$\Theta(n/\log n)$	$\Theta(n/\log n)$	$\Theta(n)$
$T_{\text{MNB}}: \Theta(n)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$

In the best case of Table 2, the time per iteration is bounded irrespectively of the dimension n , offering hope that with special computing and communication hardware, some extremely large practical problems can be solved in reasonable time.

Another interesting case, which is not covered by Table 2, arises in connection with the linear iteration $x := Ax + b$, where A is an $n \times n$ fully dense matrix. It can be shown that this iteration can be executed in a hypercube network of n^2 processors in $\Theta(\log n)$ time [see, for example, Bertsekas and Tsitsiklis (1989b)]. While it is hard to imagine at present hypercubes of n^2 processors solving large $n \times n$ systems, this case provides a theoretical limit for the time per iteration for unstructured linear systems in message-passing machines.

Consider now the possibility of using $p \ll n$ processors, each updating an n/p -dimensional component of x . The computation time per iteration will then increase by a factor n/p , so the question arises whether it is possible to improve the order of growth of the communication time in the cases where T_{MNB} is the iteration time bottleneck. The cases of medium and large T_{MNB} are of principal interest here. In these cases the corresponding times measured in message transmission time units are $\Theta(p/\log p)$ and $\Theta(p)$, respectively. Because, however, each message involves n/p values, its transmission time grows linearly with n/p , so the corresponding time T_{MNB} becomes $\Theta(n/\log p)$ and $\Theta(n)$, respectively. Thus the order of time per iteration is not improved by choosing $p \ll n$, at least under the hypotheses of this section.

5. ASYNCHRONOUS ITERATIONS

Asynchronous iterations have been introduced by Chazan and Miranker (1969) (under the name *chaotic relaxation*) for the solution of linear equations. In an asynchronous implementation of the iteration $x := f(x)$, processors are not required to wait to receive all messages generated during the previous iteration. Rather, each processor is allowed to keep iterating on its own component at its own pace. If the current value of the component updated by some other processor is not available, then some outdated value received at some time in the past is used instead. Furthermore, processors are not required to communicate their results after each iteration but only once in a while. We allow some processors to compute faster and execute more iterations than others, we allow some processors to communicate more frequently than others, and we allow the communication delays to be substantial and unpredictable. We also allow the communication channels to deliver messages out of order, i.e. in a different order than the one they were transmitted.

There are several potential advantages that may be gained from asynchronous execution [see Kung (1976) for a related discussion].

(a) *Reduction of the synchronization penalty.* There is no overhead such as the one associated with the global synchronization method. In particular, a

processor can proceed with the next iteration without waiting for all other processors to complete the current iteration, and without waiting for a synchronization algorithm to execute. Furthermore, in certain cases, there are even advantages over the local synchronization method as we now discuss. Suppose that an algorithm happens to be such that each iteration leaves the value of x_i unchanged. With local synchronization, processor i must still send messages to every processor j with $(i, j) \in A$ because processor j will not otherwise proceed to the next iteration. Consider now a somewhat more realistic case where the algorithm is such that a typical iteration is very likely to leave x_i unchanged. Then each processor j with $(i, j) \in A$ will be often found in a situation where it waits for rather uninformative messages stating that the value of x_i has not changed. In an asynchronous execution, processor j does not wait for messages from processor i and the progress of the algorithm is likely to be faster. A similar argument can be made for the case where x_i changes only slightly between iterations. Notice that the situation is similar to the case of synchronization via rollback, except that in an asynchronous algorithm processors do not roll back even if they iterate on the basis of outdated and later invalidated information.

(b) *Ease of restarting.* Suppose that the processors are engaged in the solution of an optimization problem and that suddenly one of the parameters of the problem changes. (Such a situation is common and natural in the context of data networks or in the quasistatic control of large scale systems.) In a synchronous execution, all processors should be informed, abort the computation, and then reinitiate (in a synchronized manner) the algorithm. In an asynchronous implementation no such reinitialization is required. Rather, each processor incorporates the new parameter value in its iterations as soon as it learns the new value, without waiting for all processors to become aware of the parameter change. When all processors learn the new parameter value, the algorithm becomes the correct (asynchronous) iteration.

(c) *Reduction of the effect of bottlenecks.* Suppose that the computational power of processor i suddenly deteriorates drastically. In a synchronous execution the entire algorithm would be slowed down. In an asynchronous execution, however, only the progress of x_i and of the components strongly influenced by x_i would be affected; the remaining components would still retain the capacity of making unhampered progress. Thus the effects of temporary malfunctions tend to be localized. The same argument applies to the case where a particular communication channel is suddenly slowed down.

(d) *Convergence acceleration due to a Gauss-Seidel effect.* With a Gauss-Seidel execution, convergence often takes place with fewer updates of each component, the reason being that new information is incorporated faster in the update formulas. On the other hand Gauss-Seidel iterations are generally less parallelizable. Asynchronous algorithms have the

potential of displaying a Gauss-Seidel effect because newest information is incorporated into the computations as soon as it becomes available, while retaining maximal parallelism as in Jacobi-type algorithms.

A major potential drawback of asynchronous algorithms is that they cannot be described mathematically by the iteration $x(t+1) = f(x(t))$. Thus, even if this iteration is convergent, the corresponding asynchronous iteration could be divergent, and indeed this is sometimes the case. Even if the convergence of the asynchronous iteration can be established, the corresponding analysis is often difficult. Nevertheless, there is a large number of results stating that certain classes of important algorithms retain their desirable convergence properties in the face of asynchronism: they will be surveyed in Sections 6–8. Another difficulty relates to the fact that an asynchronous algorithm may have converged (within a desired accuracy) but the algorithm does not terminate because no processor is aware of this fact. We address this issue in Section 9.

We now present our model of asynchronous computation. Let the set X and the function f be as described in Section 2. Let t be an integer variable used to index the events of interest in the computing system. Although t will be referred to as a time variable, it may have little relation with "real time". Let $x_i(t)$ be the value of x_i residing in the memory of the i th processor at time t . We assume that there is a set of times T' at which x_i is updated. To account for the possibility that the i th processor may not have access to the most recent values of the components of x , we assume that

$$x_i(t+1) = f_i(x_1(\tau'_1(t)), \dots, x_n(\tau'_n(t))), \quad \forall t \in T', \quad (5.1)$$

where $\tau'_j(t)$ are times satisfying

$$0 \leq \tau'_j(t) \leq t, \quad \forall t \geq 0.$$

At all times $t \notin T'$, $x_i(t)$ is left unchanged and

$$x_i(t+1) = x_i(t), \quad \forall t \notin T'. \quad (5.2)$$

We assume that the algorithm is initialized with some $x(0) \in X$.

The above mathematical description can be used as a model of asynchronous iterations executed by either a message-passing distributed system or a shared-memory parallel computer. For an illustration of the latter case, see Fig. 7.

The difference $t - \tau'_j(t)$ is equal to zero for a synchronous execution. The larger this difference is, the larger is the amount of asynchronism in the algorithm. Of course, for the algorithm to make any progress at all we should not allow $\tau'_j(t)$ to remain forever small. Furthermore, no processor should be allowed to drop out of the computation and stop iterating. For this reason, certain assumptions need to be imposed. There are two different types of assumptions which we state below.

Assumption 1. (Total asynchronism). The sets T' are infinite and if $\{t_k\}$ is a sequence of elements of T' which tends to infinity, then $\lim \tau'_j(t_k) = \infty$ for every j .

Assumption 2. (Partial asynchronism). There exists a positive constant B such that:

(a) For every $t \geq 0$ and every i , at least one of the elements of the set $\{t, t+1, \dots, t+B-1\}$ belongs to T' .

(b) There holds

$$t-B \leq \tau'_j(t) \leq t, \quad \forall i, j, \forall t \in T'. \quad (5.3)$$

(c) There holds $\tau'_j(t) = t$, for all i and $t \in T'$.

The constant B of Assumption 2, to be called the *asynchronism measure*, bounds the amount by which the information available to a processor can be outdated. Notice that a Jacobi-type synchronous iteration is the special case of partial asynchronism in which $B = 1$. Notice also that Assumption (c) states that the information available to processor i regarding its own component is never outdated. Such an assumption is natural in most contexts, but could be violated in certain types of shared memory parallel computing systems if we allow more than one processor to update the same component of x . It turns

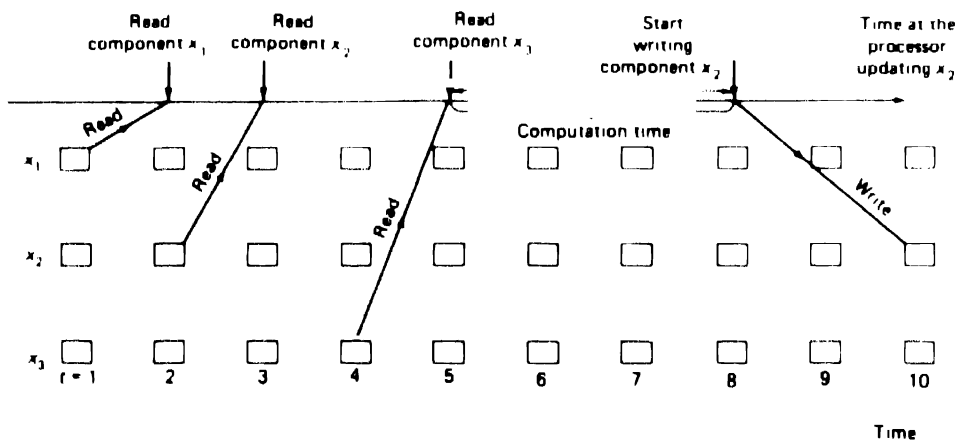


FIG. 7. Illustration of a component update in a shared memory multiprocessor. Here x_2 is viewed as being updated at time $t = 9$ ($9 \in T'$), with $\tau'_1(9) = 1$, $\tau'_2(9) = 2$, and $\tau'_3(9) = 4$. The updated value of x_2 is entered at the corresponding register at $t = 10$. Several components can be simultaneously in the process of being updated, and the values of $\tau'_j(t)$ can be unpredictable.

out that if we relax Assumption 2(c), the convergence of certain asynchronous algorithms is destroyed (Lubachevsky and Mitra, 1986; Bertsekas and Tsitsiklis, 1989b, p. 506 and p. 517). Parts (a) and (b) of Assumption 2 are typically satisfied in practice.

Asynchronous algorithms can exhibit three different types of behavior (other than guaranteed divergence).

- (a) Convergence under total asynchronism.
- (b) Convergence under partial asynchronism, for every value of B , but possible divergence under totally asynchronous execution.
- (c) Convergence under partial asynchronism if P is small enough, and possible divergence if B is large enough.

The mechanisms by which convergence is established in each one of the above three cases are fundamentally different and we address them in the subsequent three sections, respectively.

6. TOTALLY ASYNCHRONOUS ALGORITHMS

Totally asynchronous convergence results have been obtained† by Chazan and Miranker (1969) for linear iterations, Miellou (1975a), Baudet (1978), El Tarazi (1982), Miellou and Spiteri (1985) for contracting iterations, Miellou (1975b) and Bertsekas (1982) for monotone iterations, and Bertsekas (1983) for general iterations. Related results can be also found in Uresin and Dubois (1986, 1988, 1990). The following general result is from Bertsekas (1983).

Proposition 3. Let $X = \prod_{i=1}^n X_i \subset \prod_{i=1}^n \mathcal{R}^n$. Suppose that for each $i \in \{1, \dots, p\}$, there exists a sequence $\{X_i(k)\}$ of subsets of X_i such that

- (a) $X_i(k+1) \subset X_i(k)$, for all $k \geq 0$.
- (b) The sets $X(k) = \prod_{i=1}^p X_i(k)$ have the property $f(x) \in X(k+1)$, for all $x \in X$.
- (c) Every limit point of a sequence $\{x(k)\}$ with the property $x(k) \in X(k)$ for all k , is a fixed point of f .

Then, under Assumption 1 (total asynchronism), and if $x(0) \in X(0)$, every limit point of the sequence $\{x(t)\}$ generated by the asynchronous iteration (5.1)–(5.2) is a fixed point of f .

Proof. We show by induction that for each $k \geq 0$, there is a time t_k such that:

- (a) $x(t) \in X(k)$ for all $t \geq t_k$.
- (b) For all i and $t \in T^i$ with $t \geq t_k$, we have $x^i(t) \in X(k)$, where

$$x^i(t) = (x_1(\tau_1^i(t)), x_2(\tau_2^i(t)), \dots, x_n(\tau_n^i(t))), \quad \forall t \in T^i.$$

[In words: after some time, all solution estimates will be in $X(k)$ and all estimates used in iteration (5.1) will come from $X(k)$.]

† Actually, some of these papers only consider partially asynchronous iterations, but their convergence results readily extend to cover the case of total asynchronism.

The induction hypothesis is true for $k = 0$, since the initial estimate is assumed to be in $X(0)$. Assuming it is true for a given k , we will show that there exists a time t_{k+1} with the required properties. For each $i = 1, \dots, n$, let t^i be the first element of T^i such that $t^i \geq t_k$. Then by condition (b) in the statement of the proposition, we have $f(x^i(t^i)) \in X(k+1)$ and

$$x_i(t^i + 1) = f_i(x^i(t^i)) \in X_i(k+1)$$

Similarly, for every $t \in T^i$, $t \geq t^i$, we have $x_i(t+1) \in X_i(k+1)$. Between elements of T^i , $x_i(t)$ does not change. Thus,

$$x_i(t) \in X_i(k+1), \quad \forall t \geq t^i + 1.$$

Let $t_k = \max\{t^i\} + 1$. Then, using the Cartesian product structure of $X(k)$ we have

$$x(t) \in X(k+1), \quad \forall t \geq t_k$$

Finally, since by Assumption 1, we have $\tau_j^i(t) \rightarrow \infty$ as $t \rightarrow \infty$, $t \in T^i$, we can choose a time $t_{k+1} \geq t_k$ that is sufficiently large so that $\tau_j^i(t) \geq t_k^j$ for all i, j and $t \in T^i$ with $t \geq t_{k+1}$. We then have, $x_j(\tau_j^i(t)) \in X_j(k+1)$, for all $t \in T^i$ with $t \geq t_{k+1}$ and all $j = 1, \dots, n$, which implies that

$$x^i(t) = (x_1(\tau_1^i(t)), x_2(\tau_2^i(t)), \dots, x_n(\tau_n^i(t))) \in X(k+1).$$

The induction is complete. \square E.D.

The key idea behind Proposition 3 is that eventually $x(t)$ enters and stays in the set $X(k)$, furthermore, due to condition (b) in Proposition 3, it eventually moves into the next set $X(k+1)$. The most restrictive assumption in the proposition is the requirement that each $X(k)$ is the Cartesian product of sets $X_i(k)$. Successful application of Proposition 3 depends on the ability to properly define the sets $X_i(k)$ with the required properties. This is possible for two general classes of iterations which will be discussed shortly.

Notice that Proposition 3 makes no assumptions on the nature of the sets $X_i(k)$. For this reason, it can be applied to problems involving continuous variables, as well as discrete iterations involving finite-valued variables. Furthermore, the result extends in the obvious way to the case where each $X_i(k)$ is a subset of an infinite-dimensional space (instead of being a subset of \mathcal{R}^n) or to the case where f has multiple fixed points.

Interestingly enough, the sufficient conditions for asynchronous convergence provided by Proposition 3, are also known to be necessary for two special cases: (i) if $n_i = 1$ for each i and the mapping f is linear (Chazan and Miranker, 1969), and (ii) if the set X is finite (Uresin and Dubois, 1990).

Several authors have also studied asynchronous iterations with zero delays, that is, under the assumption $\tau_j^i(t) = t$ for every $t \in T^i$; see for example Robert *et al.* (1975); Robert (1976, 1987, 1988). Note that this is a special case of our asynchronous model, but is more general than the synchronous Jacobi and Gauss-Seidel iterations of Section 2, because the sets T^i are allowed to be arbitrary. General necessary and sufficient convergence conditions for the zero-delay

case can be found in Tsitsiklis (1987) where it is shown that asynchronous convergence is guaranteed if and only if there exists a Lyapunov-type function which testifies to this.

6.1. Maximum norm contractions

Consider a norm on \mathcal{R}^n defined by

$$\|x\| = \max_i \frac{\|x_i\|_i}{w_i}, \quad (6.1)$$

where $x_i \in \mathcal{R}^{n_i}$ is the i th component of x , $\|\cdot\|_i$ is a norm on \mathcal{R}^{n_i} , and w_i is a positive scalar, for each i . Suppose that f has the following contraction property: there exists some $\alpha \in [0, 1)$ such that

$$\|f(x) - x^*\| \leq \alpha \|x - x^*\|, \quad \forall x \in X, \quad (6.2)$$

where x^* is a fixed point of f . Given a vector $x(0) \in X$ with which the algorithm is initialized, let

$$X_i(k) = \{x_i \in \mathcal{R}^{n_i} \mid \|x_i - x_i^*\| \leq \alpha^k \|x(0) - x^*\|\}.$$

It is easily verified that these sets satisfy the conditions of Proposition 3 and convergence to x^* follows.

Iteration mappings f with the contraction property (6.2) are very common. We list a few examples:

(a) Linear iterations of the form $f(x) = Ax + b$, where A is an $n \times n$ matrix such that $\rho(|A|) < 1$. Here, $|A|$ is the matrix whose entries are the absolute values of the corresponding entries of A , and $\rho(|A|)$, the spectral radius of $|A|$, is the largest of the magnitudes of the eigenvalues of $|A|$ (Chazan and Miranker, 1969). This result follows from a corollary of the Perron-Frobenius theorem that states that $\rho(|A|) < 1$ if and only if A is a contraction mapping with respect to a weighted maximum norm of the form (6.1), for a suitable choice of the weights. As a special case, we obtain totally asynchronous convergence of the iteration $\pi := \pi P$ for computing a row vector π consisting of the invariant probabilities of an irreducible, discrete-time, finite-state, Markov chain. Here, P is the transition probability matrix of the chain and one of the components of π is held fixed throughout the algorithm (Bertsekas and Tsitsiklis, 1989b). Another special case, the case of periodic asynchronous iterations, is considered in Donnelly (1971). Let us mention here that the condition $\rho(|A|) < 1$ is not only sufficient but also necessary for totally asynchronous convergence (Chazan and Miranker, 1969).

(b) Gradient iterations of the form $f(x) = x - \gamma \nabla F(x)$, where γ is a small positive stepsize parameter, $F: \mathcal{R}^n \rightarrow \mathcal{R}$ is a twice continuously differentiable cost function whose Hessian matrix is bounded and satisfies the diagonal dominance condition

$$\sum_{j \neq i} |\nabla_{ij}^2 F(x)| \leq \nabla_{ii}^2 F(x) - \beta, \quad \forall i, \forall x \in X. \quad (6.3)$$

Here, β is a positive constant and $\nabla_{ij}^2 F$ stands for $(\partial^2 F)/(\partial x_i \partial x_j)$ (Bertsekas, 1983; Bertsekas and Tsitsiklis, 1989b).

Example 1. Consider the iteration $x := x - \gamma Ax$,

where A is the positive definite matrix given by

$$\begin{bmatrix} 1+\epsilon & 1 & 1 \\ 1 & 1+\epsilon & 1 \\ 1 & 1 & 1+\epsilon \end{bmatrix}$$

and γ, ϵ are positive constants. This iteration can be viewed as the gradient iteration $x := x - \gamma \nabla F(x)$ for minimizing the quadratic function $F(x) = \frac{1}{2} x' A x$ and is known to converge synchronously if the stepsize γ is sufficiently small. If $\epsilon > 1$, then the diagonal dominance condition of (6.3) holds and totally asynchronous convergence follows, when the stepsize γ is sufficiently small. On the other hand, when $0 < \epsilon < 1$, the condition of (6.3) fails to hold for all $\gamma > 0$. In fact, in that case, it is easily shown that $\rho(I - \gamma A) > 1$ for every $\gamma > 0$, and totally asynchronous convergence fails to hold, according to the necessary conditions quoted earlier. An illustrative sequence of events under which the algorithm diverges is the following. Suppose that the processors start with a common vector $x(0) = (c, c, c)$ and that each processor executes a very large number t_0 of updates of its own component without informing the others. Then, in effect, processor 1 solves the equation $0 = (\partial F / \partial x_1)(x_1, c, c) = (1 + \epsilon)x_1 + c + c$, to obtain $x_1(t_0) \approx -2c/(1 + \epsilon)$, and the same conclusion is obtained for the other processors as well. Assume now that the processors exchange their results at time t_0 and repeat the above described scenario. We will then obtain $x_i(2t_0) \approx -2x_i(t_0)/(1 + \epsilon) \approx (-2)^2 c/(1 + \epsilon)^2$. Such a sequence of events can be repeated *ad infinitum*, and it is clear that the vector $x(t)$ will diverge if $\epsilon < 1$.

(c) The projection algorithm (as well as several other algorithms) for variational inequalities. Here, $X = \bigcap_{i=1}^I X_i \subset \mathcal{R}^n$ is a closed convex set, $f: X \rightarrow \mathcal{R}^n$ is a given function, and we are looking for a vector $x^* \in X$ such that

$$(x - x^*)' f(x^*) \geq 0, \quad \forall x \in X.$$

The projection algorithm is given by $x := [x - \gamma f(x)]'$, where $[\cdot]'$ denotes orthogonal projection on the set X . Totally asynchronous convergence to x^* is obtained under the assumption that the mapping $x \mapsto x - \gamma f(x)$ is a maximum norm contraction mapping, and this is always the case if the Jacobian of f satisfies a diagonal dominance condition (Bertsekas and Tsitsiklis, 1989b). Special cases of variational inequalities include constrained convex optimization, solution of systems of nonlinear equations, traffic equilibrium problems under a user-optimization principle, and Nash games. Let us point out here that an asynchronous algorithm for solving a traffic equilibrium problem can be viewed as a model of a traffic network in operation whereby individual users optimize their individual routes given the current condition of the network. It is natural to assume that such user-optimization takes place asynchronously. Similarly, in a game theoretic context, we can think of a set of players who asynchronously adapt their strategies so as to improve their individual payoffs.

and an asynchronous iteration can be used as a model of such a situation.

(d) Waveform relaxation methods for solving a system of ordinary differential equations under a weak coupling assumption (Mitra, 1987), as well as for two-point boundary value problems (Lang *et al.*, 1986; Spiteri, 1984; Bertsekas and Tsitsiklis, 1989b).

Other studies have dealt with an asynchronous Newton algorithm (Bojanczyk, 1984), an agreement problem (Li and Basar, 1987), diagonally dominant linear programming problems (Tseng, 1990), and a variety of infinite-dimensional problems such as partial differential equations, and variational inequalities (Spiteri, 1984, 1986; Miellou and Spiteri, 1985; Anwar and El Tarazi, 1985).

In the case of maximum norm contraction mappings, there are some convergence rate estimates available which indicate that the asynchronous iteration converges faster than its synchronous counterpart, especially if the coupling between the different components of x is relatively weak. Let us suppose that an update by a processor takes one time unit and that the communication delays are always equal to D time units, where D is a positive integer. With a synchronous algorithm, there is one iteration every $D + 1$ time units and the "error" $\|x(t) - x^*\|$ can be bounded by $C\alpha^{(t/D+1)}$, where C is some constant [depending on $x(0)$] and α is the contraction factor of (6.2). We now consider an asynchronous execution whereby, at each time step, an iteration is performed by each processor i and the result is immediately transmitted to the other processors. Thus the values of x_j ($j \neq i$) which are used by processor i are always outdated by D time units. Concerning the function f , we assume that there exists some scalar β such that $0 < \beta < \alpha$ and

$$\|f_i(x) - x_i^*\| \leq \max \{ \alpha \|x_i - x_i^*\|, \beta \max_{j \neq i} \|x_j - x_j^*\| \} \quad \forall i \quad (6.4)$$

It is seen that a small value of β corresponds to a situation where the coupling between different components of x is weak. Under condition (6.4), the convergence rate estimate for the synchronous iteration cannot be improved, but the error $\|x(t) - x^*\|$ for the asynchronous iteration can be shown (Bertsekas and Tsitsiklis, 1989b) to be bounded above by $C\rho^t$, where C is some constant and ρ is the positive solution of the equation $\rho = \max \{ \alpha, \beta\rho^{-D} \}$. It is not hard to see that $\rho < \alpha^{(1/D+1)}$ and the asynchronous algorithm converges faster. The advantage of the asynchronous algorithm is more pronounced when β is very small (very weak coupling) in which case ρ approaches α . The latter is the convergence rate that would have been obtained if there were no communication delays at all. We conclude that, for weakly coupled problems, asynchronous iterations are slowed down very little by communication delays, in sharp contrast with their synchronous counterparts.

6.2. Monotone mappings

Consider a function $f: \mathcal{R}^n \rightarrow \mathcal{R}^n$ which is continuous, monotone [that is, if $x \leq y$ then $f(x) \leq f(y)$], and has a unique fixed point x^* . Furthermore, assume that there exist vectors u, v , such that $u \leq f(u) \leq f(v) \leq v$. If we let f^k be the composition of k copies of f and $X(k) = \{x \mid f^k(u) \leq x \leq f^k(v)\}$, then Proposition 3 applies and establishes totally asynchronous convergence. The above stated conditions on f are satisfied by the iteration mapping corresponding to the successive approximation (value iteration) algorithm for discounted and certain undiscounted infinite horizon dynamic programming problems (Bertsekas, 1982).

An important special case is the asynchronous Bellman-Ford algorithm for the shortest path problem. Here we are given a directed graph $G = (N, A)$, with $N = \{1, \dots, n\}$ and for each arc $(i, j) \in A$, a weight a_{ij} representing its length. The problem is to compute the shortest distance x_i from every node i to node 1. We assume that every cycle not containing node 1 has positive length and that there exists at least one path from every node to node 1. Then, the shortest distances correspond to the unique fixed point of the monotone mapping $f: \mathcal{R}^n \rightarrow \mathcal{R}^n$ defined by $f_i(x) = 0$ and

$$f_j(x) = \min_{(i,j) \in A} (a_{ij} + x_i), \quad j \neq 1$$

The Bellman-Ford algorithm consists of the iteration $x := f(x)$ and can be shown to converge asynchronously (Tajbnapis, 1977; Bertsekas, 1982). We now compare the synchronous and the asynchronous versions. We assume that both versions are initialized with $x_i = \infty$ for every $i \neq 1$, which is the most common choice. The synchronous iteration is known to converge after at most n iterations. However, assuming that the communication delays from processor i to j are fixed to some constant D_{ij} , and that the computation time is negligible, it is easily shown that the asynchronous iteration is guaranteed to terminate earlier than the synchronous one.

Notice that the number of messages exchanged in the synchronous Bellman-Ford algorithm is at most n^2 . This is because there are at most n stages and at most n messages are transmitted by each processor at each stage. Interestingly enough, with an asynchronous execution, and if the communication delays are allowed to be arbitrary, some simple examples (due to E. M. Gafni and R. G. Gallager; see Bertsekas and Tsitsiklis, 1989b) show that the number of messages exchanged until termination could be exponential in n , even if we restrict processor i to transmit a message only when the value of x_i changes. This could be a serious drawback but experience with the algorithm indicates that this worst case behavior rarely occurs and that the average number of messages exchanged is polynomial in n . It also turns out that the expected number of messages is polynomial in n under some reasonable probabilistic assumptions on the execution of the algorithm (Tsitsiklis and Stamoulis, 1990).

A number of asynchronous convergence results

making essential use of monotonicity conditions are also available for relaxation and primal-dual algorithms for linear and nonlinear network flow problems (Bertsekas, 1986; Bertsekas and Eckstein, 1987, 1988; Bertsekas and El Baz, 1987; Bertsekas and Castanon, 1989, 1990). Experiments showing faster convergence for asynchronous over synchronous relaxation methods for assignment problems using a shared memory machine are given in Bertsekas and Castanon (1989).

We finally note that, under the monotonicity assumptions of this subsection, the convergence rate of an asynchronous iteration is guaranteed to be at least as good as the convergence rate of a corresponding synchronous iteration, under a fair comparison (Bertsekas and Tsitsiklis, 1989a).

7 PARTIALLY ASYNCHRONOUS ALGORITHMS—I

We now consider iterations satisfying the partial asynchronism Assumption 2. Since old information is "purged" from the algorithm after at most B units, it is natural to describe the "state" of the algorithm at time t by the vector $z(t) \in X^n$ defined by

$$z(t) = (x(t), x(t-1), \dots, x(t-B+1)).$$

We then notice that $x(t+1)$ can be determined [cf. (5.1)–(5.3)] in terms of $z(t)$, in particular, knowledge of $x(\tau)$, for $\tau \leq t-B$ is not needed. We assume that the iteration mapping f is continuous and has a nonempty set $X^* \subset X$ of fixed points. Let Z^* be the set of all vectors $z^* \in X^n$ of the form $z^* = (x^*, x^*, \dots, x^*)$, where x^* belongs to X^* . We present a sometimes useful convergence result, which employs a Lyapunov-type function d defined on the set X^n .

Proposition 4. (Bertsekas and Tsitsiklis, 1989b) Suppose that there exist a positive integer r^* and a continuous function $d: X^n \rightarrow [0, \infty)$ with the following properties: For every initialization $z(0) \notin Z^*$ of the iteration and any subsequent sequence of events (conforming to Assumption 2) we have $d(z(r^*)) \leq d(z(0))$ and $d(z(1)) \leq d(z(0))$. Then every limit point of a sequence $\{z(t)\}$ generated by the partially asynchronous iteration (5.1)–(5.2) belongs to Z^* . Furthermore, if $X = \mathcal{H}^n$, if the function d is of the form $d(z) = \inf_{z^* \in Z^*} \|z - z^*\|$, where $\|\cdot\|$ is some vector

norm, and if the function f is of the form $f(x) = Ax + b$, where A is a $n \times n$ matrix and b is a vector in \mathcal{H}^n , then $d(z(t))$ converges to zero at the rate of a geometric progression.

For an interesting application of the above proposition, consider a mapping $f: \mathcal{H}^n \rightarrow \mathcal{H}^n$ of the form $f(x) = Ax$ where A is an irreducible stochastic matrix, and let $n_i = 1$ for each i . In the corresponding iterative algorithm, each processor maintains and communicates a value of a scalar variable x_i and once in a while forms a convex combination of its own variable with the variables received from other processors according to the rule

$$x_i := \sum_{j=1}^n a_{ij} x_j.$$

Clearly, if the algorithm converges then, in the limit, the values possessed by different processors are equal. We will thus refer to the asynchronous iteration $x := Ax$ as an *agreement* algorithm. It can be shown that, under the assumption of partial asynchronism, the function d defined by

$$d(z(t)) = \max_{i \in \{1, \dots, n\}} \max_{\tau \in \{t-B+1, \dots, t\}} x_i(\tau) - \min_{i \in \{1, \dots, n\}} \min_{\tau \in \{t-B+1, \dots, t\}} x_i(\tau) \quad (7.1)$$

has the properties assumed in Proposition 4, provided that at least one of the diagonal entries of A is positive. In particular, if the processors initially disagree, the "maximum disagreement" [cf. (7.1)] is reduced by a positive amount after at most $2nB$ time units (Tsitsiklis, 1984). Proposition 4 applies and establishes geometric convergence to agreement. Furthermore, such partially asynchronous convergence is obtained no matter how big the value of the asynchronism measure B is, as long as B is finite.

The following example (Bertsekas and Tsitsiklis, 1989b) shows that the agreement algorithm need not converge totally asynchronously.

Example 2. Suppose that

$$A = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}.$$

Here, the synchronous iteration $x(t+1) = Ax(t)$ converges in a single step to the vector $x = (y, y)$, where $y = (x_1 + x_2)/2$. Consider the following totally asynchronous scenario. Each processor updates its value at each time step. At certain times t_1, t_2, \dots , each processor transmits its value which is received with zero delay and is immediately incorporated into the computations of the other processor. We then have

$$x_1(t+1) = \frac{x_1(t)}{2} + \frac{x_2(t_k)}{2}, \quad t_k \leq t < t_{k+1},$$

$$x_2(t+1) = \frac{x_2(t_k)}{2} + \frac{x_1(t)}{2}, \quad t_k \leq t < t_{k+1}.$$

(See Fig. 8 for an illustration.) Thus,

$$x_1(t_{k+1}) = (1/2)^{t_{k+1}-t_k} x_1(t_k) + (1 - (1/2)^{t_{k+1}-t_k}) x_2(t_k),$$

$$x_2(t_{k+1}) = (1/2)^{t_{k+1}-t_k} x_2(t_k) + (1 - (1/2)^{t_{k+1}-t_k}) x_1(t_k).$$

Subtracting these two equations we obtain

$$\begin{aligned} |x_2(t_{k+1}) - x_1(t_{k+1})| &= (1 - 2(1/2)^{t_{k+1}-t_k}) |x_2(t_k) - x_1(t_k)| \\ &= (1 - \epsilon_k) |x_2(t_k) - x_1(t_k)|, \end{aligned}$$

where $\epsilon_k = 2(1/2)^{t_{k+1}-t_k}$. In particular, the disagreement $|x_2(t_k) - x_1(t_k)|$ keeps decreasing. On the other hand, convergence to agreement is not guaranteed unless $\prod_{k=1}^{\infty} (1 - \epsilon_k) = 0$ which is not necessarily the case. For example, if we choose the differences $t_{k+1} - t_k$ to be large enough so that $\epsilon_k < k^{-2}$, then we can use the fact $\prod_{k=1}^{\infty} (1 - k^{-2}) > 0$ to see that convergence to agreement does not take place.

Example 2 shows that failure to converge is possible if part (b) of the partial asynchronism Assumption 2

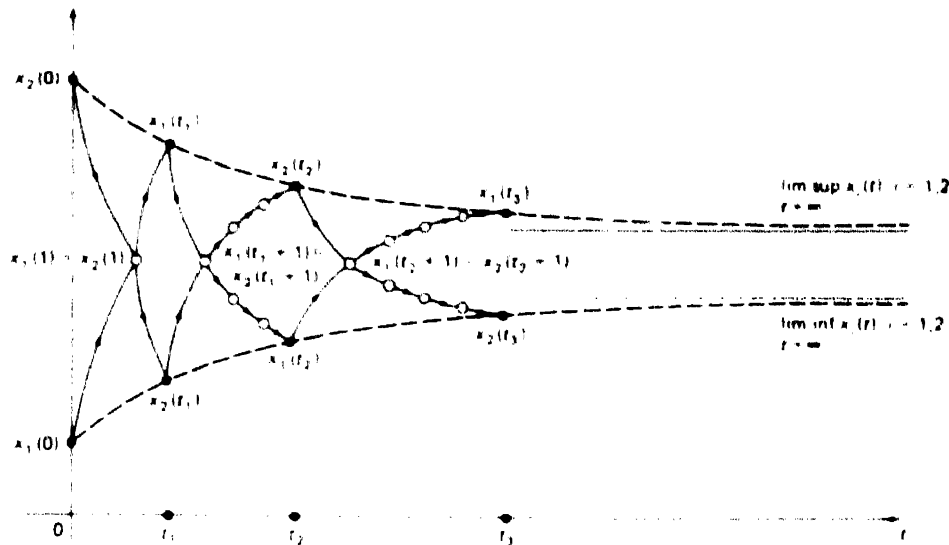


FIG. 8. Illustration of divergence in Example 2.

facts to hold. There also exist examples demonstrating that parts (a) and (c) of Assumption 2 are also necessary for convergence.

Example 2 illustrates best the convergence mechanism in algorithms which converge partially asynchronously for every B , but not totally asynchronously. The key idea is that the distance from the set of fixed points is guaranteed to "contract" once in a while. However, the contraction factor depends on B and approaches 1 as B gets larger. (In the context of Example 2, the contraction factor is $1 - \epsilon_k$ which approaches 1 as $t_{k+1} - t_k$ is increased to infinity.) As time goes to infinity, the distance from the set of fixed points is contracted an infinite number of times but this guarantees convergence only if the contraction factor is bounded away from 1, which then necessitates a finite but otherwise arbitrary bound on B .

Partially asynchronous convergence for every value of B has been established for several variations and generalizations of the agreement algorithm (Tsitsiklis, 1984; Bertsekas and Tsitsiklis, 1989b), as well as for a variety of other problems:

(a) The iteration $\pi := \pi P$ for the computation of a row vector π of invariant probabilities, associated with an irreducible stochastic matrix P with a nonzero diagonal entry (Lubachevsky and Mitra, 1986). This result can be also obtained by letting $x_i = \pi_i / \pi_i^*$, where π^* is a positive vector satisfying $\pi^* = \pi^* P$, and by verifying that the variables x_i obey the equations of the agreement algorithm (Bertsekas and Tsitsiklis, 1989b).

(b) Relaxation algorithms involving nonexpansive mappings with respect to the maximum norm (Tseng *et al.*, 1990; Bertsekas and Tsitsiklis, 1989b). Special cases include dual relaxation algorithms for strictly convex network flow problems and linear iterations for the solution of linear equations of the form $Ax = b$, where A is an irreducible matrix satisfying the weak diagonal dominance condition $\sum_{j \neq i} |a_{ij}| \leq a_{ii}$, for all i .

(c) An asynchronous algorithm for load balancing in a computer network whereby highly loaded processors transfer fractions of their load to their lightly loaded neighbors, until the load of all processors becomes the same (Bertsekas and Tsitsiklis, 1989b).

In all of the above cases, partially asynchronous convergence has been proved for all values of B , and examples are available which demonstrate that totally asynchronous convergence fails.

We close by mentioning a particular context in which the agreement algorithm could be of use. Consider a set of processors who obtain a sequence of noisy observations and try to estimate certain parameters by means of some iterative method. This could be a stochastic gradient algorithm (such as the ones arising in recursive system identification) or some kind of a Monte Carlo estimation algorithm. All processors are employed for the estimation of the same parameters but their individual estimates are generally different because the noises corrupting their observations can be different. We let the processors communicate and combine their individual estimates in order to average their individual noises, thereby reducing the error variance. We thus let the processors execute the agreement algorithm, trying to agree on a common estimate, while simultaneously obtaining new observations which they incorporate into their estimates. There are two opposing effects here: the agreement algorithm tends to bring their estimates closer together, while new observations have the potential of increasing the difference of their estimates. Under the partial asynchronism assumption, the agreement algorithm tends to converge geometrically. On the other hand, in several stochastic algorithms (such as the stochastic approximation iteration

$$x := x - \frac{1}{t} (\nabla F(x) + w),$$

where w represents observation noise) the stepsize $1/t$

decreases to zero as time goes to infinity. We then have, asymptotically, a separation of time scales: the stochastic algorithm operates on a slower time scale and therefore the agreement algorithm can be approximated by an algorithm in which agreement is instantly established. It follows that the asynchronous nature of the agreement algorithm cannot have any adverse effect on the convergence of the stochastic algorithm. Rigorous results of this type can be found in Tsitsiklis (1984), Tsitsiklis *et al.* (1986); Kushner and Yin (1987a, b); Bertsekas and Tsitsiklis (1989b).

8. PARTIALLY ASYNCHRONOUS ALGORITHMS—II

We now turn to the study of partially asynchronous iterations that converge only when the stepsize is small. We illustrate the behavior of such algorithms in terms of a prototypical example.

Let A be an $n \times n$ positive definite symmetric matrix and let b be a vector in \mathcal{R}^n . We consider the asynchronous iteration $x := x - \gamma(Ax - b)$, where γ is a small positive stepsize. We define a cost function $F: \mathcal{R}^n \rightarrow \mathcal{R}$ by $F(x) = \frac{1}{2}x'Ax - x'b$, and our iteration is equivalent to the gradient algorithm $x := x - \gamma \nabla F(x)$ for minimizing F . This algorithm is known to converge synchronously provided that γ is chosen small enough. On the other hand, it was shown in Example 1 that the gradient algorithm does not converge totally asynchronously. Furthermore, a careful examination of the argument in that example reveals that for every value of γ there exists a B large enough such that the partially asynchronous gradient algorithm does not converge. Nevertheless, if γ is fixed to a small value, and if B is not excessively large (we roughly need $B < C/\gamma$, where C is some constant determined by the structure of the matrix A), then the partially asynchronous iteration turns out to be convergent. An equivalent statement is that for every value of B there exists some $\gamma_0 > 0$ such that if $0 < \gamma < \gamma_0$ then the partially asynchronous algorithm converges (Tsitsiklis *et al.*, 1986; Bertsekas and Tsitsiklis, 1989b). The rationale behind such a result is the following. If the information available to processor i on the value of x is outdated by at most B time units, then the difference between the value $x_i(\tau_i(t))$ possessed by processor i and the true value $x_i(t)$ is of the order of γB , because each step taken by processor j is of the order of γ . It follows that for γ very small the errors caused by asynchronism become negligible and cannot destroy the convergence of the algorithm.

The above mentioned convergence result can be extended to more general gradient-like algorithms for nonquadratic cost functions F . One only needs to assume that the iteration is of the form $x := x - \gamma s(x)$, where $s(x)$ is an update direction with the property $s_i(x) \nabla_i F(x) \geq K \|\nabla_i F(x)\|^2$, where K is a positive constant, together with a Lipschitz continuity condition on ∇F , and a boundedness assumption of the form $\|s(x)\| \leq L \|\nabla F(x)\|$ (Tsitsiklis *et al.*, 1986; Bertsekas and Tsitsiklis, 1989b). Similar conclusions are obtained for gradient projection iterations for constrained convex optimization (Bertsekas and

Tsitsiklis, 1989b).

An important application of asynchronous gradient-like optimization algorithms arises in the context of optimal quasistatic routing in data networks. In a common formulation of the routing problem one is faced with a convex nonlinear multicommodity network flow problem (Bertsekas and Gallager, 1987) that can be solved using gradient projection methods. It has been shown that these methods also converge partially asynchronously, provided that a small enough stepsize is used (Tsitsiklis and Bertsekas, 1986). Furthermore, such methods can be naturally implemented on-line by having the processors in the network asynchronously exchange information on the current traffic conditions in the system and perform updates trying to reduce the measure of congestion being optimized. An important property of such an asynchronous algorithm is that it adapts to changes in the problem being solved (such as changes on the amount of traffic to be routed through the network) without a need for aborting and restarting the algorithm. Some further analysis of the asynchronous routing algorithm can be found in Tsai (1986, 1989) and Tsai *et al.* (1986).

9. TERMINATION OF ASYNCHRONOUS ITERATIONS

In practice, iterative algorithms are executed only for a finite number of iterations, until some termination condition is satisfied. In the case of asynchronous iterations, the problem of determining whether termination conditions are satisfied is rather difficult because each processor possesses only partial information on the progress of the algorithm.

We now introduce one possible approach for handling the termination problem for asynchronous iterations. In this approach, the problem is decomposed into two parts:

- (a) An asynchronous iterative algorithm is modified so that it terminates in finite time.
- (b) A special procedure is used to detect termination in finite time after it has occurred.

In order to handle the termination problem, we have to be a little more specific about the model of interprocessor communication. While the general model of asynchronous iterations introduced in Section 5 can be used for both shared memory and message-passing parallel architectures, we adopt here a more explicit message-passing model. In particular, we assume that each processor j sends messages with the value of x_j to every other processor i . Processor i keeps a buffer with the most recently received value of x_j . We denote the value in this buffer at time t by $x_j'(t)$. This value was transmitted by processor j at some earlier time $\tau_j'(t)$ and therefore $x_j'(t) = x_j(\tau_j'(t))$. We also assume the following:

Assumption 3. (a) If $t \in T'$ and $x_i(t+1) \neq x_i(t)$, then processor i will eventually send a message to every other processor.

(b) If a processor i has sent a message with the

value of $x_i(t)$ to some other processor j , then processor i will send a new message to processor j only after the value of x_i changes (due to an update by processor i).

(c) Messages are received in the order that they are transmitted.

(d) Each processor sends at least one message to every other processor.

Assumption 3(d) is only needed to get the algorithm started. Assumption 3(b) is crucial and has the following consequences. If the value of $x(t)$ settles to some final value, then there will be some time t^* after which no messages will be sent. Furthermore, all messages transmitted before t^* will eventually reach their destinations and the algorithm will eventually reach a quiescent state where none of the variables x_i changes and no message is in transit. We can then say that the algorithm has terminated.

More formally, we view termination as equivalent to the following two properties:

- (i) No message is in transit
- (ii) An update by some processor i causes no change in the value of x_i .

Property (ii) is a collection of local termination conditions. There are several algorithms for termination detection when a termination condition can be decomposed as above (Dijkstra and Scholten, 1980; Bertsekas and Tsitsiklis, 1989b). Thus termination detection causes no essential difficulties, under the assumption that the asynchronous algorithm terminates in finite time.

We now turn to the more difficult problem of converting a convergent asynchronous iterative algorithm into a finitely terminating one. If we were dealing with the synchronous iteration $x(t+1) = f(x(t))$, it would be natural to terminate the algorithm when the condition $\|x(t+1) - x(t)\| \leq \epsilon$ is satisfied, where ϵ is a small positive constant reflecting the desired accuracy of solution, and where $\|\cdot\|$ is a suitable norm. This suggests the following approach for the context of asynchronous iterations. Given the iteration mapping f and the accuracy parameter ϵ , we define a new iteration mapping $g: X \rightarrow X$ by letting

$$g_i(x) = \begin{cases} f_i(x) & \text{if } \|f_i(x) - x_i\| \leq \epsilon, \\ x_i, & \text{otherwise.} \end{cases}$$

We will henceforth assume that the processors are executing the asynchronous iteration $x := g(x)$. The key question is whether this new iteration is guaranteed to terminate in finite time. One could argue as follows. Assuming that the original iteration $x := f(x)$ is guaranteed to converge, the changes in the vector x will eventually become arbitrarily small, in which case we will have $g(x) = x$ and the iteration $x := g(x)$ will terminate. Unfortunately, this argument is fallacious, as demonstrated by the following example.

Example 3. Consider the function $f: \mathcal{R}^2 \rightarrow \mathcal{R}^2$

defined by

$$f_1(x) = \begin{cases} -x_1, & \text{if } x_2 \geq \epsilon/2, \\ 0, & \text{if } x_2 < \epsilon/2, \end{cases}$$

$$f_2(x) = x_2/2.$$

It is clear that the asynchronous iteration $x := f(x)$ converges to $x^* = (0, 0)$; in particular, x_2 is updated according to $x_2 := x_2/2$ and tends to zero; thus, it eventually becomes smaller than $\epsilon/2$. Eventually processor 1 receives a value of x_2 smaller than $\epsilon/2$ and a subsequent update by the same processor sets x_1 to zero.

Let us now consider the iteration $x := g(x)$. If the algorithm is initialized with x_2 between $\epsilon/2$ and ϵ , then the value of x_2 will never change, and processor 1 will keep executing the nonconvergent iteration $x_1 := -x_1$. Thus, the asynchronous iteration $x := g(x)$ is not guaranteed to terminate.

The remainder of this section is devoted to the derivation of conditions under which the iteration $x := g(x)$ is guaranteed to terminate. We introduce some notation. Let I be a subset of the set $\{1, \dots, p\}$ of all processors. For each $i \in I$, let there be given some value $\theta_i \in X_i$. We consider the asynchronous iteration $x := f'^n(x)$, which is the same as the iteration $x := f(x)$ except that any component x_i , with $i \in I$, is set to the value θ_i . Formally, the mapping f'^n is defined by letting $f'_i(x) = f_i(x)$, if $i \notin I$, and $f'_i(x) = \theta_i$, if $i \in I$.

Proposition 5. (Bertsekas and Tsitsiklis, 1989a) Let Assumption 3 hold. Suppose that for any $I \subseteq \{1, \dots, p\}$ and for any choice of $\theta_i \in X_i$, $i \in I$, the asynchronous iteration $x := f'^n(x)$ is guaranteed to converge. Then, the asynchronous iteration $x := g(x)$ terminates in finite time.

Proof. Consider the asynchronous iteration $x := g(x)$. Let I be the set of all indices i for which the variable $x_i(t)$ changes only a finite number of times. For each $i \in I$, let θ_i be the limiting value of $x_i(t)$. Since f maps X into itself, so does g . It follows that $\theta_i \in X_i$ for each i . For each $i \in I$, processor i sends a positive but finite number of messages [Assumptions 3(d) and (b)]. By Assumption 3(a), the last message sent by processor i carries the value θ_i , and by Assumption 3(c) this is also the last message received by any other processor. Thus, for all t large enough, and for all j , we will have $x'_j(t) = x_j(\tau_j(t)) = \theta_j$. Thus, the iteration $x := g(x)$ eventually becomes identical with the iteration $x := f'^n(x)$ and therefore converges. This implies that the difference $x_i(t+1) - x_i(t)$ converges to zero for any $i \notin I$. On the other hand, because of the definition of the mapping g , the difference $x_i(t+1) - x_i(t)$ is either zero, or its magnitude is bounded below by $\epsilon > 0$. It follows that $x_i(t+1) - x_i(t)$ eventually settles to zero, for every $i \notin I$. This shows that $i \in I$ for every $i \notin I$; we thus obtain a contradiction unless $I = \{1, \dots, p\}$, which proves the desired result. Q.E.D.

We now identify certain cases in which the main assumption in Proposition 5 is guaranteed to hold. We consider first the case of monotone iterations and we

assume that the iteration mapping f has the properties introduced in Section 6.2. For any I and $\{\theta_i | i \in I\}$, the mapping $f^{I,n}$ inherits all of the properties of f , except that $f^{I,n}$ is not guaranteed to have a unique fixed point. If this latter property can be independently verified, then the asynchronous iteration $x := f^{I,n}(x)$ is guaranteed to converge, and Proposition 3 applies. Let us simply say here that this property can be indeed verified for several interesting problems.

Let us now consider the case where f satisfies the contraction condition $\|f(x) - x^*\| \leq \alpha \|x - x^*\|$ of (6.2). Unfortunately, it is not necessarily true that the mappings $f^{I,n}$ also satisfy the same contraction condition. In fact, the mappings $f^{I,n}$ are not even guaranteed to have a fixed point. Let us strengthen the contraction condition of (6.2) and assume that

$$\|f(x) - f(y)\| \leq \alpha \|x - y\|, \quad \forall x, y \in \mathcal{R}^n, \quad (9.1)$$

where $\|\cdot\|$ is the weighted maximum norm of (6.1) and $\alpha \in [0, 1)$. We have $f^{I,n}(x) - f^{I,n}(y) = \theta_i - \theta_i = 0$ for all $i \in I$. Thus,

$$\begin{aligned} \|f^{I,n}(x) - f^{I,n}(y)\| &= \max_{i \in I} \frac{1}{w_i} \|f_i(x) - f_i(y)\|, \\ &\leq \max_{i \in I} \frac{1}{w_i} \|f_i(x) - f_i(y)\|, \\ &= \|f(x) - f(y)\| \leq \alpha \|x - y\|. \end{aligned}$$

Hence, the mappings $f^{I,n}$ inherit the contraction property (9.1). As discussed in Section 6, this property guarantees asynchronous convergence and therefore Proposition 5 applies again.

We conclude that the modification $x := g(x)$ of the asynchronous iteration $x := f(x)$ is often, but not always, guaranteed to terminate in finite time. It is an interesting research question to devise economical termination procedures for the iteration $x := f(x)$ that are always guaranteed to work. The snapshot algorithm of Chandy and Lamport (1985) [see (Bertsekas and Tsitsiklis, 1989, Section 8.2)] seems to be one option.

10. CONCLUSIONS

Iterative algorithms are easy to parallelize and can be executed synchronously even in inherently asynchronous computing systems. Furthermore, for the regular communication networks associated with several common parallel architectures, the communication requirements of iterative algorithms are not severe enough to preclude the possibility of massive parallelization and speedup of the computation. Iterative algorithms can also be executed asynchronously, often without losing the desirable convergence properties of their synchronous counterparts, although the mechanisms that affect convergence can be quite different for different types of algorithms. Such asynchronous execution may offer substantial advantages in a variety of contexts.

At present, there is very strong evidence suggesting that asynchronous iterations converge faster than their synchronous counterparts. However, this evidence is principally based on analysis and simulations. There is only a small number of related experimental works using shared memory machines. These works support

the conclusions of the analysis but more testing with a broader variety of computer architectures is needed to provide a comprehensive picture of the practical behavior of asynchronous iterations. Furthermore, the proper implementation of asynchronous algorithms in real parallel machines can be quite challenging and more experience is needed in this area. Finally, much remains to be done to enlarge the already substantial class of problems for which asynchronous algorithms can be correctly applied.

REFERENCES

- Anwar, M. N. and N. El Tarazi (1985). Asynchronous algorithms for Poisson's equation with nonlinear boundary conditions. *Computing*, **34**, 155-168.
- Awerbuch, B. (1985). Complexity of network synchronization. *J. ACM*, **32**, 804-823.
- Barbosa, V. C. (1986). Concurrency in systems with neighborhood constraints. Doctoral Dissertation, Computer Science Dept., U.C.L.A., Los Angeles, CA, U.S.A.
- Barbosa, V. C. and E. M. Gafni (1987). Concurrency in heavily loaded neighborhood-constrained systems. *Proc. 7th Int. Conf. on Distributed Computing Systems*.
- Baudet, G. M. (1978). Asynchronous iterative methods for multiprocessors. *J. ACM*, **2**, 226-244.
- Bertsekas, D. P. (1982). Distributed dynamic programming. *IEEE Trans. Aut. Control*, **AC-27**, 610-616.
- Bertsekas, D. P. (1983). Distributed asynchronous computation of fixed points. *Math. Programm.*, **27**, 107-120.
- Bertsekas, D. P. (1986). Distributed asynchronous relaxation methods for linear network flow problems. Technical Report LIDS-P-1606, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA.
- Bertsekas, D. P. and D. A. Castanon (1989). Parallel synchronous and asynchronous implementations of the auction algorithm. Technical Report TP-308, Alphatech Inc., Burlington, MA. Also *Parallel Computing* (to appear).
- Bertsekas, D. P. and D. A. Castanon (1990). Parallel asynchronous primal-dual methods for the minimum cost flow problem. *Math. Programming* (submitted).
- Bertsekas, D. P. and J. Eckstein (1988). Dual coordinate step methods for linear network flow problems. *Math. Programming*, **42**, 203-243.
- Bertsekas, D. P. and J. Eckstein (1987). Distributed asynchronous relaxation methods for linear network flow problems. *Proc. IFAC '87*, Munich, F.R.G.
- Bertsekas, D. P. and D. El Baz (1987). Distributed asynchronous relaxation methods for convex network flow problems. *SIAM J. Control Optimiz.*, **25**, 74-84.
- Bertsekas, D. P. and R. G. Gallager (1987). *Data Networks*. Prentice Hall, Englewood Cliffs, NJ.
- Bertsekas, D. P., C. Ozveren, G. Stamoulis, P. Tseng and J. N. Tsitsiklis (1989). Optimal communication algorithms for hypercubes. Technical Report LIDS-P-1847, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA. Also *J. Parallel Distrib. Comput.* (to appear).
- Bertsekas, D. P. and J. N. Tsitsiklis (1989a). Convergence Rate and Termination of Asynchronous Iterative Algorithms. *Proc. 1989 Int. Conf. on Supercomputing*, Iraklion, Greece, 1989, pp. 461-470.
- Bertsekas, D. P. and J. N. Tsitsiklis (1989b). *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ.
- Bojanczyk, A. (1984). Optimal asynchronous Newton method for the solution of nonlinear equations. *J. ACM*, **32**, 792-803.
- Chandy, K. M. and L. Lamport (1985). Distributed snapshots: determining global states of distributed systems. *ACM Trans. Comput. Syst.*, **3**, 63-75.
- Chazan, D. and W. Miranker (1969). Chaotic relaxation. *Linear Algebra and its Applications*, **2**, 199-222.
- Donnelly, J. D. P. (1971). Periodic chaotic relaxation. *Linear Algebra and its Applications*, **4**, 117-128.

- Dijkstra, E. W. and C. S. Scholten (1980). Termination detection for diffusing computations. *Inform. Process. Lett.*, **11**, 1-4.
- Dubois, M. and F. A. Briggs (1982). Performance of synchronized iterative processes in multi-processor systems. *IEEE Trans. Software Engng.*, **8**, 419-431.
- El Tarazi, M. N. (1982). Some convergence results for asynchronous algorithms. *Numerische Mathematik*, **39**, 325-340.
- Fox, G., M. Johnson, G. Lyzenga, S. Otto, J. Salmon and D. Walker (1988). *Solving Problems on Concurrent Processors, Vol. 1*. Prentice-Hall, Englewood Cliffs, NJ.
- Jefferson, D. R. (1985). Virtual time. *ACM Trans. Programming Languages Syst.*, **7**, 404-425.
- Johansson, S. L. and C. T. Ho (1989). Optimum broadcasting and personalized communication in hypercubes. *IEEE Trans. Comput.*, **38**, 1249-1268.
- Kung, H. T. (1976). Synchronized and asynchronous parallel algorithms for multiprocessors. In *Algorithms and Complexity*. Academic Press, New York, pp. 153-200.
- Kushner, H. J. and G. Yin (1987a). Stochastic approximation algorithms for parallel and distributed processing. *Stochastics*, **22**, 219-250.
- Kushner, H. J. and G. Yin (1987b). Asymptotic properties of distributed and communicating stochastic approximation algorithms. *SIAM J. Control Optimiz.*, **25**, 1266-1290.
- Lang, B., J. C. Miellou and P. Spiteri (1986). Asynchronous relaxation algorithms for optimal control problems. *Main Comput. Simult.*, **28**, 227-242.
- Lavenberg, S., R. Muntz and B. Samadi (1983). Performance analysis of a rollback method for distributed simulation. In A. K. Agrawala and S. K. Tripathi (Eds.), *Performance '83*. North Holland, Amsterdam, pp. 117-132.
- Li, S. and T. Basar (1987). Asymptotic agreement and convergence of asynchronous stochastic algorithms. *IEEE Trans. Aut. Control*, **32**, 612-618.
- Lubachevsky, B. and D. Mitra (1986). A chaotic asynchronous algorithm for computing the fixed point of a non-negative matrix of unit spectral radius. *J. ACM*, **33**, 130-150.
- McBryan, O. A. and E. F. Van der Velde (1987). Hypercube algorithms and implementations. *SIAM J. Scientific Statist. Comput.*, **8**, s227-s287.
- Miellou, J. C. (1975a). Algorithmes de relaxation chaotique a retards. *R.A.I.R.O.*, **9**, R.1, 55-82.
- Miellou, J. C. (1975b). Iterations chaotiques a retards: etudes de la convergence dans le cas d'espaces partiellement ordonnes. *Comptes Rendus, Academie de Sciences de Paris*, **280**, Serie A, 233-236.
- Miellou, J. C. and P. Spiteri (1985). Un critere de convergence pour des methodes generales de point fixe. *Math. Modelling Numer. Anal.*, **19**, 645-669.
- Mitra, D. (1987). Asynchronous relaxations for the numerical solution of differential equations by parallel processors. *SIAM J. Sci. Statist. Comput.*, **8**, s43-s58.
- Mitra, D. and I. Mitran (1984). Analysis and optimum performance of two message-passing parallel processors synchronized by rollback. In E. Gelenbe (Ed.), *Performance '84*. North Holland, Amsterdam, 35-50.
- Nassimi, D. and S. Sahni (1980). An optimal routing algorithm for mesh-connected parallel computers. *J. ACM*, **27**, 6-29.
- Ortega, J. M. and R. G. Voigt (1985). Solution of partial differential equations on vector and parallel computers. *SIAM Review*, **27**, 149-240.
- Ozveren, C. (1987). Communication aspects of parallel processing. Technical Report LIDS-P-1721, Laboratory for Information and Decision Systems, MIT, Cambridge, MA.
- Robert, F. (1976). Contraction en norme vectorielle, convergence d'iterations chaotiques pour des equations non lineaires de point fixe a plusieurs variables. *Linear Algebra and its Applications*, **13**, 19-35.
- Robert, F. (1987). Iterations discretees asynchrones. Technical Report 671M, I.M.A.G., University of Grenoble, France.
- Robert, F., M. Charnay and F. Musy (1975). Iterations chaotiques serie-parallele pour des equations non-lineaires de point fixe. *Aplikace Matematiky*, **20**, 1-38.
- Saad, Y. and M. H. Schultz (1987). Data Communication in Hypercubes. Research Report YALEU/DCS/RR-428, Yale University, New Haven, CN.
- Smart, D. and J. White (1988). Reducing the parallel solution time of sparse circuit matrices using reordered Gaussian elimination and relaxation. *Proc. 1988 ISCAS*, Espoo, Finland.
- Spiteri, P. (1984). Contribution a l'etude de grands systemes non lineaires. Doctoral Dissertation, l'Universite de Franche-Comte, Besancon, France.
- Spiteri, P. (1986). Parallel asynchronous algorithms for solving boundary value problems. In M. Cosnard et al. (Eds.), *Parallel Algorithms and Architectures*. North Holland, Amsterdam, pp. 73-84.
- Tajbnapis, W. D. (1977). A correctness proof of a topology information maintenance protocol for a distributed computer network. *Commun. ACM*, **20**, 477-485.
- Tsai, W. K. (1986). Optimal quasi-static routing for virtual circuit networks subjected to stochastic inputs. Doctoral Dissertation, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA.
- Tsai, W. K. (1989). Convergence of gradient projection routing methods in an asynchronous stochastic quasi-static virtual circuit network. *IEEE Trans. Aut. Control*, **34**, 20-33.
- Tsai, W. K., J. N. Tsitsiklis and D. P. Bertsekas (1986). Some issues in distributed asynchronous routing in virtual circuit data networks. *Proc. 25th IEEE Conf. on Decision and Control*, Athens, Greece, 1335-1337.
- Tseng, P. (1990). Distributed computation for linear programming problems satisfying a certain diagonal dominance condition. *Math. Operat. Res.*, **15**, 33-48.
- Tseng, P., D. P. Bertsekas and J. N. Tsitsiklis (1990). *SIAM J. Control Optimiz.*, **28**, 678-710.
- Tsitsiklis, J. N. (1984). Problems in decentralized decision making and computation. Ph.D. thesis, Dep. of Electrical Engineering and Computer Science, MIT, Cambridge, MA.
- Tsitsiklis, J. N. (1987). On the stability of asynchronous iterative processes. *Math. Syst.*, **20**, 137-153.
- Tsitsiklis, J. N. (1989). A comparison of Jacobi and Gauss-Seidel parallel iterations. *Applied Math. Lett.*, **2**, 167-170.
- Tsitsiklis, J. N. and D. P. Bertsekas (1986). Distributed asynchronous optimal routing in data networks. *IEEE Trans. Aut. Control*, **AC-31**, 325-332.
- Tsitsiklis, J. N., D. P. Bertsekas and M. Athans (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Trans. Aut. Control*, **AC-31**, 803-812.
- Tsitsiklis, J. N. and G. D. Stamoulis (1990). On the average communication complexity of asynchronous distributed algorithms. Technical Report LIDS-P-1986, Laboratory for Information and Decision Systems, MIT, Cambridge, MA.
- Uresin, A. and M. Dubois (1986). Generalized asynchronous iterations. In *Lecture Notes in Computer Science*, **237**, Springer, Berlin, pp. 272-278.
- Uresin, A. and M. Dubois (1988a). Sufficient conditions for the convergence of asynchronous iterations. Technical Report, Computer Research Institute, University of Southern California, Los Angeles, California, U.S.A.
- Uresin, A. and M. Dubois (1990). Parallel asynchronous algorithms for discrete data. *J. ACM*, **37**, 588-606.
- Zemos, S. A. and R. A. Lasken (1988). Nonlinear network optimization on a massively parallel Connection Machine. *Annals of Operations Research*, **14**, 147-165.
- Zemos, S. A. and J. M. Mulvey (1988). Distributed algorithms for strictly convex network flow problems. *Parallel Computing*, **6**, 45-56.

A Frequency-domain Estimator for Use in Adaptive Control Systems*

RICHARD O. LAMAIRES†, LENA VALAVANI§||, MICHAEL ATHANS§
and GUNTER STEIN§†

A robust estimation technique, developed for adaptive control systems, finds both a parametrized model and a corresponding frequency-domain error bounding function.

Key Words—Adaptive control; identification; frequency domain; estimation; parameter estimation; transfer functions; robust control

Abstract—This paper presents a frequency-domain estimator that can identify both a parametrized nominal model of a plant as well as a frequency-domain bounding function on the modeling error associated with this nominal model. This estimator, which we call a robust estimator, can be used in conjunction with a robust control-law redesign algorithm to form a robust adaptive controller.

1. INTRODUCTION AND MOTIVATION

THE USE OF feedback control in systems having large amounts of uncertainty requires the use of algorithms that learn or adapt in an on-line situation. A control system that is designed using only *a priori* knowledge results in a relatively low bandwidth closed-loop system so as to guarantee stable operation in the face of large uncertainty. An adaptive control algorithm, which can identify the plant on-line, thereby decreasing the amount of uncertainty, can yield a closed-loop system that has a higher bandwidth and thus better performance than a non-adaptive algorithm. There are many problems with the adaptive control algorithms that have been

developed to date. In particular, most adaptive control algorithms are not robust to unmodeled dynamics and an unmeasurable disturbance, particularly in the absence of a persistently-exciting input signal.

In this section, we will motivate the robust estimation problem by first discussing the adaptive control problem, in general, and then presenting a perspective on the robust adaptive control problem. Further, we justify the choice of an infrequent adaptation strategy before discussing the main focus of the paper, the development of a robust estimator.

Stability of adaptive control algorithms

The use of adaptive control yields systems that are nonlinear and time-varying. Thus, the stability of these systems depends on the inputs and disturbances, as well as the plant (including any unmodeled dynamics) and the compensator. However, the stability properties of a linear time-invariant (LTI) feedback system depend only on the plant and compensator, not the inputs and disturbances. Because of this fact, we take the point of view that it is desirable to make the system “as LTI as possible”.

The preceding argument can be used to justify an infrequent control-law redesign strategy. It is envisioned that a discrete-time estimator will be used to continually update the frequency-domain estimate of the plant as long as there is useful information in the input/output data of the plant. The plant is in a closed loop that is controlled by a discrete-time compensator that is only infrequently updated (redesigned). It can be shown that if the compensator is redesigned sufficiently infrequently, then the LTI stability of the “frozen” system at every point in time guarantees the exponential stability of the

*Received 27 October 1988; revised 5 July 1989; received in final form 23 February 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor C. R. Johnson Jr under the direction of Editor P. C. Parks.

† Honeywell Systems and Research Center, Minneapolis, MN 55418, U.S.A.

‡ IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A. Author to whom all correspondence should be addressed.

§ Laboratory for Information and Decision Systems, Dept of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

|| Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

time-varying system. In this way, the control system looks nearly LTI and consequently is more robust to disturbances than a highly nonlinear adaptive controller. It is emphasized here that a robust adaptive controller that slowly learns and produces successively better LTI compensators is the end product envisioned in this paper. However, the paper aims to develop only the estimation part of this robust adaptive controller and does not provide proofs concerning the research issue of adaptive control system performance (i.e. whether or not the resulting compensators are better for control). On the other end of the adaptive control spectrum are algorithms that quickly adapt to a changing system. As mentioned earlier, these types of controllers have poor robustness properties in that they are highly sensitive to unmodeled dynamics and unmeasurable disturbances, particularly in the absence of persistent excitation.

A perspective on the robust adaptive control problem

With the solution of the adaptive control problem for the ideal case—that is, when there are no unmodeled dynamics nor unmeasurable disturbances—the problem of robustness has become a focus of current research. Recently, a new perspective on the robust adaptive control problem has appeared in the literature (Goodwin *et al.*, 1985a). Briefly, a *robust* adaptive controller is viewed as a combination of a robust estimator and a robust control law. This is an appealing point of view. For example, if the robust estimator is not getting any useful information and consequently, is not able to improve on the current knowledge of the plant, then the adaptation aspect of the algorithm can be disabled and the adaptive controller reduces to a robust control law; that is, in a situation where the adaptive algorithm is not learning, the adaptive controller becomes simply the best robust LTI control law that one could design based only on *a priori* information and any additional information learned since the algorithm began.

Brief statement of the robust estimation problem

The main focus of this paper is the development of a robust estimator for use in an adaptive controller. In non-adaptive robust control, the designer must first obtain a nominal model along with some measure of its goodness. A practical measure of goodness is a bounding function on the magnitude of the modeling errors in the frequency-domain. Since non-adaptive robust control requires these steps, the same steps must implicitly, or explicitly, be

present in a robust adaptive control scheme, the difference being that the steps are carried out on-line rather than off-line. Thus, we assume that our robust estimator must supply (1) a nominal plant model and (2) a frequency-domain bounding function on the magnitude of the modeling uncertainty between the true plant and this nominal model. So, the robust estimator must provide an estimate of the parameters for the structure of the nominal model, as well as a frequency-domain uncertainty bounding function corresponding to this nominal model. As will be clarified later, the robust estimator requires an *a priori* bound on the unstructured uncertainty associated with the structured nominal model of the plant. It is the goal of the robust estimator to reduce the structured uncertainty and to bound the total uncertainty, which is composed of both structured and unstructured uncertainty.

Given the information of 1 and 2 above, several robust control-law design methodologies could be used, including the LQG/LTR design methodology (Athans, 1986). The envisioned adaptive control system is illustrated in Fig. 1. In this paper, we will use a discrete-time model of a sampled-data control system.

The robust estimator presented in this paper differs from most other estimators in that it provides guarantees concerning the current estimate of the nominal model of the plant. This requirement is essential if the estimator is to be used in a robust adaptive control situation. If the estimator cannot provide guarantees about the model it provides to the control-law redesign algorithm, then the redesign algorithm cannot guarantee stability of the closed-loop system. We will use a deterministic framework throughout the paper, since guarantees of stability are sought.

Related literature

Recently, there has been growing interest in the use of frequency-domain methods for the

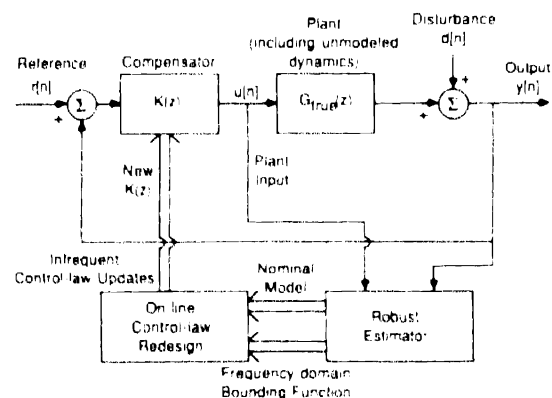


FIG. 1. A robust adaptive control system.

estimation of modeling uncertainty in the context of robust adaptive control. This work has its roots in the area of transfer function estimation. For an insightful discussion of transfer function estimation, bias distribution and many references, see Ljung (1987). Kosut (1986, 1987) has developed an approach for adaptive uncertainty modeling and on-line robust control design. In this work, the least-squares method is used to estimate the parametric model and a standard spectral estimator is used on the residual output error resulting in a crude bound on the unmodeled dynamics and disturbance spectrum. In Kosut (1988), a parameter *set* estimator is developed which provides a set of parameter values, and hence, a measure of parameter estimation accuracy. These results are similar to those reported in Morrison and Walker (1988). Both our work and Kosut's make use of the frequency-domain estimation work of Ljung (1985, 1987). Ljung analyzed the properties of the *empirical transfer function estimate* (ETFE), which is computed using the Fourier transforms of finite-length input/output data of the plant. In Ljung (1987), a constant bound on the effects of using finite-length data to compute the ETFE is developed, for strictly stable plants. This work provided the background for our development in Section 4.1 of a time-varying frequency-domain error bounding function that is computed using the DFTs of the plant input signal

In addition to the above work, Parker and Bitmead (1987a, b) have developed a promising method for estimating input/output spectra using Kalman filtering techniques, instead of discrete Fourier transforms. They also present a technique for approximating the resulting transfer function estimate with a high-order finite

impulse response (FIR) filter. In Parker and Bitmead (1987b), a bound on the L^2 norm of the model error is developed. Unfortunately, this bound does not appear to be computable on-line and it does not provide any frequency specific information about the model error.

The primary contribution of the work described in our paper is the development of bounds on the frequency-domain estimation error. These bounds can be computed on-line (at great computational expense in some cases) thereby enabling the on-line design of a robust control law. We have not yet been able to prove that the robust estimator described in this paper yields a nominal model that converges to the true plant. However, the performance of the robust estimator has been investigated, in a closed-loop adaptive control context, through simulation (e.g. see Section 8). Based on our simulation results, we believe that the robust estimator shows promise compared with current adaptive control schemes. The work that is described in this paper was originally presented in LaMaire (1987) and LaMaire *et al.* (1987).

2. MATHEMATICAL PRELIMINARIES

In this section, we will present the notation and definitions that will be used in the paper, as well as some results and theorems that will be useful later on. We denote a discrete-time signal by $x[n] = x(nT)$ where $x(t)$ denotes the sampled continuous-time signal and where n is an integer and T the sampling period. We denote the z -transform and the discrete-time Fourier transform (DTFT) of $x[n]$ by $X(z)$ and $X(e^{j\omega T})$, respectively (see Oppenheim *et al.*, 1983). Further, we denote the N -point discrete Fourier transform (DFT) of the N -point sequence $x[n]$, $n = 0, \dots, N-1$, by $X_N(\omega_k)$ where $\omega_k = (k/N)\omega_s$ for $k = 0, \dots, N-1$ and where $\omega_s = 2\pi/T$ is the sampling frequency. Since we will not always be working with N -point sequences that begin at 0, we define the following version of the DFT for a sequence of N points ending with time index n :

$$X_N^n(\omega_k) = \sum_{m=0}^n x[m] W_N^{km},$$

$$\text{for } k = 0, \dots, N-1, \quad (2.1)$$

where

$$W_N = e^{-j(2\pi/N)} \quad (2.2)$$

Signal processing theorems

In this subsection, we will develop results that can be used to bound the effects of using finite-length data to compute frequency-domain

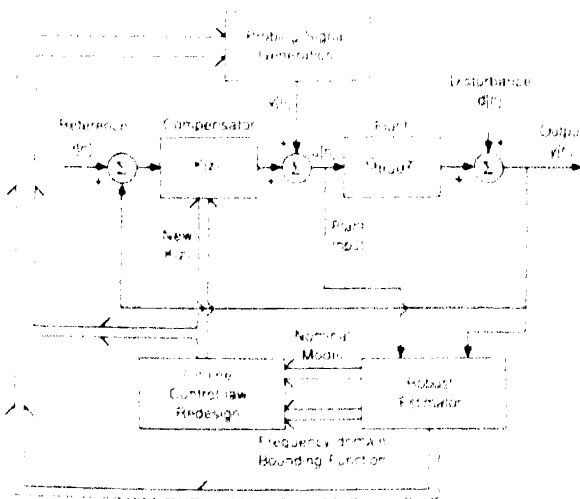


Fig. 2. A robust adaptive control system with probing signal

quantities. In the later parts of this paper, the frequency-domain estimate of a stable, causal, transfer function $H(e^{j\omega T})$ will be computed based on the N -point DFTs of the transfer function's input and output signals. We will now state a theorem that bounds the error in the frequency domain between this DFT derived frequency-domain estimate and the true transfer function.

Theorem 1. Let $y[m] = h[m] * u[m]$, where $h[m]$ is an infinite-length, causal, impulse response with all its poles in the open unit disk. We denote the DTFT of $h[m]$ by $H(e^{j\omega T})$, and the DFTs of the N -points of $u[m]$ and $y[m]$ ending with time index n , by $U_N^n(\omega_k)$ and $Y_N^n(\omega_k)$, respectively. Then,

$$Y_N^n(\omega_k) = H(e^{j\omega_k T})U_N^n(\omega_k) + E_N^n(\omega_k), \quad \text{for } k = 0, \dots, N-1, \quad (2.3)$$

where the discrete function $E_N^n(\omega_k)$ is given by

$$E_N^n(\omega_k) = \sum_{p=1}^n h[p] W_N^{kp} (U_N^{n-p}(\omega_k) - U_N^n(\omega_k)), \quad \text{for } k = 0, \dots, N-1, \quad (2.4)$$

and where W_N is defined in equation (2.2).

Proof. See Appendix.

Remark 1. The function $E_N^n(\omega_k)$ is the error in the frequency domain, at time index n , due to the use of finite-length data. That is, if the DTFTs (based on infinite-length data) of $u[m]$ and $y[m]$ were used in equation (2.3) instead of the DFTs (based on finite-length data), then there would be no error term $E_N^n(\omega_k)$. Note that the function $E_N^n(\omega_k)/U_N^n(\omega_k)$ is the error in the frequency domain between the DFT derived frequency-domain estimate of $H(e^{j\omega_k T})$ and the true transfer function $H(e^{j\omega_k T})$.

It will later be useful to be able to find a magnitude bounding function on $E_N^n(\omega_k)$. The following theorem provides such a bounding function by using only a finite summation of the DFT differences and therefore can be implemented in practice.

Theorem 2. Under the assumptions of Theorem 1 we find that given some finite integer M , the magnitude of $E_N^n(\omega_k)$ is bounded for each k as follows,

$$|E_N^n(\omega_k)| \leq \sum_{p=1}^{M-1} |h[p]| |U_N^{n-p}(\omega_k) - U_N^n(\omega_k)| + 2u_{\max} \sum_{p=M}^n p |h[p]|, \quad \text{for } k = 0, \dots, N-1, \quad (2.5)$$

where $u_{\max} = \sup |u[m]|$.

Proof. See Appendix.

3. ROBUST ESTIMATOR PROBLEM STATEMENT

In this section, we first list the assumptions required by the robust estimator and then we state the robust estimation problem. Consider the system of Fig. 1 where the discrete-time plant $G_{\text{true}}(z)$ has an input $u[n]$ and an output $y[n]$ that is corrupted by an additive output disturbance $d[n]$.

(A1) Plant assumptions

We assume a structure for the nominal model of $G_{\text{true}}(z)$ and a magnitude bounding function on the unstructured uncertainty. That is, we assume that

$$G_{\text{true}}(z) = G(z, \theta_0)[1 + \delta_u(z)] \quad (3.1)$$

where $G(z, \theta_0)$ is a nominal model, $\delta_u(z)$ denotes the unstructured uncertainty of the plant, θ_0 is a vector of plant parameters and we assume

$$(A1.1) \quad G(z, \theta_0) = \frac{B(z)}{A(z)}, \quad (3.2)$$

where the polynomials $B(z)$ and $A(z)$ are

$$B(z) = b_0 z^{(m_1-n_1)} + b_1 z^{(m_1-n_1-1)} + \dots + b_{m_1} z, \quad (3.3)$$

$$A(z) = 1 - a_1 z^{-1} - \dots - a_{m_1} z^{-m_1}, \quad (3.4)$$

and where the parameter vector is

$$\theta_0 = [a_1 \dots a_{m_1} b_0 b_1 \dots b_{m_1}]^T. \quad (3.5)$$

(A1.2) $\theta_0 \in \Theta$, where Θ is a known bounded set. (3.6)

$$(A1.3) \quad |\delta_u(e^{j\omega T})| \leq \Delta_u(e^{j\omega T}), \quad \forall \omega. \quad (3.7)$$

$$(A1.4) \quad \frac{d\delta_u(e^{j\omega T})}{d\omega} \leq \nabla_u(e^{j\omega T}), \quad \forall \omega. \quad (3.8)$$

(A1.5) $G_{\text{true}}(z)$ and $G(z, \theta_0)$ have all their poles in the open unit disk, for all $\theta_0 \in \Theta$.

(A1.6) A coarse bounding function on the magnitude of the impulse response of the true plant, denoted by $g_{\text{true}}[n]$, is known such that

$$|g_{\text{true}}[n]| \leq \sum_{i=1}^{I_0} g_i n^{(r_i)} p_i^n, \quad (3.9)$$

where r_i is a positive integer, and $g_i > 0$, $0 < p_i < 1$ (i.e. all the poles of $g_{\text{true}}[n]$ are in the open unit disk), and r_i are known for $i = 1, \dots, I_0$. $g_{\text{true}}[n]$ is assumed to be causal.

(A1.7) Zero initial conditions.

Thus, our *a priori* assumptions are that we know m_1 and n_1 , the degrees of $B(z)$ and $A(z)$, respectively, and the bounding functions $\Delta_u(e^{j\omega T})$ and $\nabla_u(e^{j\omega T})$. Further, we assume that the parameter vector θ_0 is in some known bounded set Θ which is only a coarse, and hence large, *a priori* estimate of the parameter space. While there are many ways of characterizing the high-frequency unmodeled dynamics of a plant, the description of A1.3 has been shown to have practical utility and has been used extensively in the robust control field.

(A2) *Disturbance assumption*

We assume that the N -point DFT of the disturbance signal $d[n]$, whose DFT is denoted by $D_N^d(\omega_k)$, satisfies

$$|D_N^d(\omega_k)| \leq \tilde{D}_N(\omega_k), \quad \text{for } k = 0, \dots, N-1, \forall n. \quad (3.10)$$

(A3) *Input signal assumption*

We assume that the input signal $u[n]$ is bounded and that we know u_{\max} where

$$|u[n]| \leq u_{\max}, \quad \forall n. \quad (3.11)$$

Remark 2. The discrete-time system of Fig. 1 represents a sampled-data control system. While the above plant assumptions (A1) have been presented for a discrete-time system, similar assumptions can be stated for a continuous-time plant and then used to satisfy the above discrete-time assumptions. This process, including the derivation of a discrete-time unstructured uncertainty bound from a bound on the continuous-time unstructured uncertainty, is treated in LaMaire (1987).

Remark 3. Based on input/output measurements alone we cannot determine a unique θ_0 for the nominal model because of the unstructured uncertainty. That is, if we assume the structure of A1.1 above and assume only that $\delta_u(z) \in \mathbf{S}$ where

$$\mathbf{S} = \{ \delta(z) \mid |\delta(e^{j\omega T})| \leq \Delta_u(e^{j\omega T}), \forall \omega \}, \quad (3.12)$$

then we can define a smallest set

$$\begin{aligned} \Theta^* &= \{ \theta \mid G_{\text{true}}(z) \\ &= G(z, \theta)[1 + \delta_u(z)] \text{ and } \delta_u(z) \in \mathbf{S} \} \end{aligned} \quad (3.13)$$

in which θ_0 lies. Thus, $\theta_0 \in \Theta^* \subset \Theta$ where only Θ is known *a priori*. Note that, in general, Θ^* will be a point only when $\Delta_u(e^{j\omega T}) = 0$ for all ω .

Preparation for problem statement

We rewrite the true discrete-time plant of equation (3.1) as

$$G_{\text{true}}(z) = G(z, \hat{\theta})[1 + \delta_u(z, \hat{\theta})] \quad (3.14)$$

where again $G(z, \hat{\theta})$ is the nominal model using an estimate $\hat{\theta}$ of the parameter vector θ_0 in the structure of Assumption A1.1, and $\delta_u(z, \hat{\theta})$ denotes the modeling error due to both structured and unstructured uncertainty. That is, since *a priori* we only know that $\hat{\theta} \in \Theta$, where $\hat{\theta}$ is not necessarily in Θ^* , there is structured uncertainty associated with this choice of $\hat{\theta}$ as well as the ever present unstructured uncertainty.

Problem statement

The robust estimator must provide:

- (1) a parameter estimate $\hat{\theta}$, and hence a nominal model $G(z, \hat{\theta})$,
- (2) a corresponding bounding function, $\Delta_u^n(e^{j\omega T}, \hat{\theta})$, such that

$$|\delta_u(e^{j\omega T}, \hat{\theta})| \leq \Delta_u^n(e^{j\omega T}, \hat{\theta}), \quad \forall \omega. \quad (3.15)$$

That is, at a given time index n we want to generate a new nominal model $G(z, \hat{\theta})$ (where $\hat{\theta}$ is the parameter estimate at time index n) along with a corresponding bounding function $\Delta_u^n(e^{j\omega T}, \hat{\theta})$ in the frequency domain indicating how good the current nominal model is. Given 1 and 2 above and a compensator, we can use discrete-time versions of the stability-robustness tests of Lehtomaki *et al.* (1984) to guarantee stability in the face of bounded modeling uncertainty.

The goal of the robust estimator is to find a $\hat{\theta}$ in Θ^* and to have $\Delta_u^n(e^{j\omega T}, \hat{\theta})$ approach $\Delta_u(e^{j\omega T})$. The viewpoint taken here is that the unstructured uncertainty $\Delta_u(e^{j\omega T})$ is the best we can do given the structure of our nominal model. Thus, even though $\Delta_u^n(e^{j\omega T}, \hat{\theta})$ can conceivably become smaller than our *a priori* assumed bound $\Delta_u(e^{j\omega T})$ we will not let this occur and will instead view the function $\Delta_u(e^{j\omega T})$ as the desirable lower bound of the function $\Delta_u^n(e^{j\omega T}, \hat{\theta})$.

The problem that we have described in this subsection will be referred to as the robust estimation problem. An algorithm that satisfies this problem will be referred to as a robust estimator since it provides a nominal model of the plant as well as a guaranteed frequency-domain bounding function on the accuracy of this nominal model.

Outline of problem solution

We will develop a solution to the robust estimation problem stated above. First, in Section 4, we will develop a method for computing a frequency-domain estimate of the true plant along with a bounding function on the additive error in the frequency domain. Then, in Section 5, the frequency-domain estimate of

Using equations (5.2) and (5.7–8) we can write

$$P(G(e^{j\omega_k T}, \theta_0))\theta_0 = Q(G(e^{j\omega_k T}, \theta_0)). \quad (5.9)$$

In summary, we have shown how knowledge of the complex values of $G(e^{j\omega_k T}, \theta_0)$ at the $(N/2) + 1$ frequencies $\omega_0, \dots, \omega_{(N/2)}$ can be used to write $N + 2$ linear equations in the parameters. In the ideal situation where one could exactly find $G(e^{j\omega_k T}, \theta_0)$ for $k = 0, \dots, (N/2)$, the matrix equation (5.9) will have a solution. That is, given the matrices P and Q , we could solve for the true parameter vector using any m of the linear equations, where again m is the dimension of the parameter vector θ_0 . However, in practice we will only have our cumulative frequency-domain estimate $G_{cumf, N}^n(\omega_k)$ with which to estimate the parameters. If we use $G_{cumf, N}^n(\omega_k)$ instead of $G(e^{j\omega_k T}, \theta_0)$ in equations (5.7–8), then the equation

$$P(G_{cumf, N}^n(\omega_k))\hat{\theta} = Q(G_{cumf, N}^n(\omega_k)) \quad (5.10)$$

will not, in general, have a solution. Equation (5.10) is in the form of the standard least-squares problem, which is discussed in Strang (1980).

We will choose the parameter estimate $\hat{\theta}$ as the vector that minimizes the frequency weighted norm of the error vector,

$$P(G_{cumf, N}^n(\omega_k))\hat{\theta} - Q(G_{cumf, N}^n(\omega_k)). \quad (5.11)$$

We define, with reference to equations (5.7, 5.8), the diagonal frequency weighting matrix

$$W = \text{diag}[f(\omega_0) \cdots f(\omega_{(N/2)})f(\omega_0) \cdots f(\omega_{(N/2)})] \quad (5.12)$$

where $f(\cdot)$ is the frequency weighting function. The parameter estimate that minimizes the Euclidean norm of the error vector

$$W(P(G_{cumf, N}^n(\omega_k))\hat{\theta} - Q(G_{cumf, N}^n(\omega_k))) \quad (5.13)$$

is given by the well-known result

$$\hat{\theta} = (P^T W^T W P)^{-1} P^T W^T W Q \quad (5.14)$$

where the P and Q matrices in this equation depend on the values of the estimate $G_{cumf, N}^n(\omega_k)$. Note that if P is not of full-rank then we cannot solve for $\hat{\theta}$ using equation (5.14). In practice, the *a priori* parameter estimates could be used. This situation can arise when there is insufficient excitation for the parameters to be identified. Since this insufficient excitation case would result in a large frequency-domain error bounding function (at least at some frequencies) it is unlikely that the control-law would be updated.

To gain insight as to what weighting function to choose, we notice that the estimate yielded by

equation (5.14) minimizes

$$\begin{aligned} & \sum_{k=0}^{N/2} f^2(\omega_k) |A(e^{j\omega_k T})G_{cumf, N}^n(\omega_k) - B(e^{j\omega_k T})|^2 \\ &= \sum_{k=0}^{N/2} f^2(\omega_k) |A(e^{j\omega_k T})|^2 \\ & \quad G_{cumf, N}^n(\omega_k) - \frac{B(e^{j\omega_k T})^2}{A(e^{j\omega_k T})} \end{aligned} \quad (5.15)$$

where $A(z)$ and $B(z)$ are as defined in equations (3.3, 3.4). From equation (5.15) we see that if we want our parameter estimation method to be a least-squares fit in the frequency-domain, then we want to choose a weighting function that is one over the magnitude of the denominator of the nominal model. Alternatively, we could choose the following weighting function

$$f(\omega_k) = \frac{1}{|A(e^{j\omega_k T})| \hat{E}_{cumf, N}^n(\omega_k)} \quad (5.16)$$

which provides a better fit between the nominal model and the frequency-domain estimate in the frequency ranges where the uncertainty is the smallest. Of course, we do not know what the denominator of the nominal model really is, so one can only approximately choose this frequency weighting function.

Due to the complexities (e.g. equation 4.16) of how the cumulative frequency-domain estimate $G_{cumf, N}^n(\omega_k)$ is generated from the input/output data, we have not been able to prove that the parameter estimator developed in this section yields an estimated nominal model that approaches the true plant. Some insight concerning excitation conditions under which the frequency-domain estimate $G_{f, N}^n(\omega_k)$, which is closely related to $G_{cumf, N}^n(\omega_k)$, approaches the true plant can be gained from the results of pure transfer function estimation as described in Ljung (1985, 1987). The simulations of Section 8 demonstrate the performance of the parameter estimator in a closed-loop system.

6. COMPUTING A FREQUENCY-DOMAIN UNCERTAINTY BOUNDING FUNCTION

In this section, we discuss the computation of a frequency-domain uncertainty bounding function for the nominal model $G(e^{j\omega_k T}, \hat{\theta})$. Specifically, we will compute a magnitude bounding function, $\Delta_{su}^n(e^{j\omega_k T}, \hat{\theta})$, on $\delta_{su}(e^{j\omega_k T}, \hat{\theta})$ at the frequency points ω_k for $k = 0, \dots, N - 1$.

6.1. Basic methodology

The nominal model at time index n is obtained by using the nominal model structure and the

current parameter vector estimate $\hat{\theta}$ yielded by the parameter estimator described in Section 5. Thus, we can compute the value of the nominal model $G(e^{j\omega_k T}, \hat{\theta})$ for $k = 0, \dots, N-1$. Now, using the triangle inequality, we find that at time index n , and for frequency ω_k ,

$$|G(e^{j\omega_k T}, \hat{\theta}) - G_{\text{true}}(e^{j\omega_k T})| \leq |G(e^{j\omega_k T}, \hat{\theta}) - G_{\text{cumf}, N}^n(\omega_k)| + |G_{\text{cumf}, N}^n(\omega_k) - G_{\text{true}}(e^{j\omega_k T})| \quad (6.1)$$

and using equations (4.14, 4.15),

$$|G(e^{j\omega_k T}, \hat{\theta}) - G_{\text{true}}(e^{j\omega_k T})| \leq |G(e^{j\omega_k T}, \hat{\theta}) - G_{\text{cumf}, N}^n(\omega_k)| + \tilde{E}_{\text{cumf}, N}^n(\omega_k). \quad (6.2)$$

We now can find a bound on $\delta_{su}(e^{j\omega_k T}, \hat{\theta})$. Rewriting equation (3.14),

$$G_{\text{true}}(e^{j\omega_k T}) = G(e^{j\omega_k T}, \hat{\theta})[1 + \delta_{su}(e^{j\omega_k T}, \hat{\theta})], \quad \text{for } k = 0, \dots, N-1. \quad (6.3)$$

So, rearranging yields

$$\delta_{su}(e^{j\omega_k T}, \hat{\theta}) = \frac{G_{\text{true}}(e^{j\omega_k T}) - G(e^{j\omega_k T}, \hat{\theta})}{G(e^{j\omega_k T}, \hat{\theta})}. \quad (6.4)$$

Thus, using equation (6.2), we find the bounding function

$$|\delta_{su}(e^{j\omega_k T}, \hat{\theta})| \leq \Delta_{su}^n(e^{j\omega_k T}, \hat{\theta}), \quad (6.5)$$

where

$$\Delta_{su}^n(e^{j\omega_k T}, \hat{\theta}) = \frac{|G(e^{j\omega_k T}, \hat{\theta}) - G_{\text{cumf}, N}^n(\omega_k)| + \tilde{E}_{\text{cumf}, N}^n(\omega_k)}{|G(e^{j\omega_k T}, \hat{\theta})|}, \quad \text{for } k = 0, \dots, N-1 \quad (6.6)$$

and where we have included a superscript n after the Δ_{su} to denote the fact that this bound on $|\delta_{su}(e^{j\omega_k T}, \hat{\theta})|$ depends on the time index n , since $G_{\text{cumf}, N}^n(\omega_k)$, $\tilde{E}_{\text{cumf}, N}^n(\omega_k)$ and also $\hat{\theta}$ depend on n .

In summary, we have shown how to compute a discrete function $\Delta_{su}^n(e^{j\omega_k T}, \hat{\theta})$ that bounds the net effect of structured and unstructured uncertainty of the current nominal model $G(e^{j\omega_k T}, \hat{\theta})$ relative to the true plant, at the frequencies $\omega_0, \dots, \omega_{N-1}$. We used the nominal model structure of A1.1, the current parameter estimate $\hat{\theta}$; and the cumulative frequency-domain estimate $G_{\text{cumf}, N}^n(\omega_k)$ and corresponding cumulative frequency-domain error bounding function $\tilde{E}_{\text{cumf}, N}^n(\omega_k)$, which were developed in Section 4.

6.2. A smoothed uncertainty bounding function

In this subsection, we discuss the computation of a smoothed, magnitude bounding function on $|\delta_{su}|$. This development is motivated by the observation that, depending upon the spectrum of the input signal, one may have a very jagged

bounding function on the modeling uncertainty $|\delta_{su}(e^{j\omega_k T}, \hat{\theta})|$. That is, at the frequency point ω_k the bound $\Delta_{su}^n(e^{j\omega_k T}, \hat{\theta})$ may be very tight, however, at an adjacent frequency point ω_{k+1} the bound $\Delta_{su}^n(e^{j\omega_{k+1} T}, \hat{\theta})$ may be very poor. In LaMaire (1987), it is shown how the assumptions of Section 3 can be used to find a derivative bounding function $\nabla_{su}^n(e^{j\omega T})$ satisfying

$$\frac{d\delta_{su}(e^{j\omega T}, \hat{\theta})}{d\omega} \leq \nabla_{su}^n(e^{j\omega T}), \quad \forall \omega. \quad (6.7)$$

If δ_{su} is analytic, then it is shown (LaMaire, 1987) that

$$|\delta_{su}(e^{j\omega T}, \hat{\theta})| \leq |\delta_{su}(e^{j\omega_k T}, \hat{\theta})| + (\omega - \omega_k) \nabla_{su}^n(\omega_k, \omega_{k+1}) \quad (6.8)$$

and

$$|\delta_{su}(e^{j\omega T}, \hat{\theta})| \leq |\delta_{su}(e^{j\omega_{k+1} T}, \hat{\theta})| + (\omega_{k+1} - \omega) \nabla_{su}^n(\omega_k, \omega_{k+1}) \quad (6.9)$$

for $\omega \in [\omega_k, \omega_{k+1}]$ where

$$\nabla_{su}^n(\omega_k, \omega_{k+1}) = \sup_{\omega \in [\omega_k, \omega_{k+1}]} \{ \nabla_{su}^n(e^{j\omega T}) \}. \quad (6.10)$$

From these equations we see that it may be possible to obtain a tighter bound on $|\delta_{su}(e^{j\omega_k T}, \hat{\theta})|$ than $\Delta_{su}^n(e^{j\omega_k T}, \hat{\theta})$, by using the bound at an adjacent frequency point, $\Delta_{su}^n(e^{j\omega_{k+1} T}, \hat{\theta})$ or $\Delta_{su}^n(e^{j\omega_{k-1} T}, \hat{\theta})$, along with the smoothness information of $\Delta_{su, n}$.

6.3. Bounding inter-sample variations

In this brief subsection, we discuss the computation of a safety factor that must be added to the discrete bounding function $\Delta_{su}^n(e^{j\omega_k T}, \hat{\theta})$ to account for inter-sample variations. Ultimately, the uncertainty bounding function at discrete frequency points will be used in stability-robustness tests to design a new robust compensator. These stability-robustness tests are meant to be used with continuous functions of frequency. Since the actual computations will be performed with an uncertainty bounding function that is a discrete function of frequency, we must add the aforementioned safety factor to the discrete function to account for the worst possible peaks that may occur between frequency samples ω_k . In LaMaire (1987), it is shown how equations (6.8, 6.9) can be used to choose this additive safety factor in such a way that the largest inter-sample variations lie below a line drawn between the values of the final uncertainty bounding function (including the safety factor) at two adjacent frequency samples.

7. TIME-DOMAIN PARAMETER ESTIMATION: AN ALTERNATIVE

In this section, we briefly describe an alternative method to that of Section 5 for generating the parameter estimate $\hat{\theta}$ defining the nominal model $G(z, \hat{\theta})$. In LaMaire (1987), a time-domain parameter estimator was developed that adjusts the parameter estimates selectively depending upon the usefulness of the input/output data. This time-domain estimator is a combination of the modified least-squares algorithm that was developed by Goodwin *et al.* (1985b, 1986) and the bounding mechanism of Theorem A.2 in the appendix. Goodwin *et al.*'s modified least-squares algorithm is made robust through the use of a time-varying dead-zone. It is shown in LaMaire (1987) how Theorem A.2 can be used to bound the time-domain effects of high-frequency unmodeled dynamics and an unmeasurable disturbance given the assumptions of the robust estimator (i.e. the frequency-domain bound on the unmodeled dynamics and the bound on the disturbance DFT) and the on-line computed DFTs of the input signal. Unfortunately, our simulations revealed that the robust time-domain parameter estimator described in this section did not perform well due to the conservatism of the resulting time-domain bound on the unmodeled dynamics and disturbance effects. The parameter estimator was "turned-off" by the time-varying dead-zone in many situations where a standard least-squares algorithm was able to continue yielding accurate parameter estimates. Consequently, we chose to use the frequency-domain method of Section 5 to determine parameter estimates for the nominal model of the robust estimator.

8. SIMULATION EXAMPLES

In this section, we will present several simulation examples to illustrate the behavior of the robust estimator. First, in Subsection 8.1 we consider a simple first-order plant in an open-loop disturbance-free situation. This example gives insight into the basic frequency-domain bounding methodology and serves to illustrate the different types of behavior that arise for two types of input signals. In Subsection 8.2, we will describe a high-order plant model and its associated adaptive control system. The simulation results for this high-order example, which are presented in Subsection 8.3, demonstrate that if the input signal is rich, then the robust estimator does lead to improved closed-loop performance.

8.1. Basic frequency-domain bounding illustration

The z -transform and impulse response of the discrete-time first-order plant that we consider in this subsection are given by

$$H(z) = \frac{r}{z - p} \quad \text{and} \quad (8.1)$$

$$h[n] = gp^n, \quad n \geq 1, \quad (8.2)$$

respectively, where $r = 0.46651$, $p = 0.53349$ and $g = r/p = 0.87446$. This discrete-time example corresponds to the zero-order hold equivalent (with a sampling period of $\pi/5$ sec) of a unity-gain continuous-time plant that has a pole at 1 rad sec^{-1} . For this simple illustrative example we assume that we know *a priori* the exact impulse response bounding function. For this case we are examining the performance of the frequency-domain bounding methodology of equation (4.6) with $g_1 = g$ and $p_1 = p$, and for design choices $M = 10$ and $N = 50$. So, using equations (4.6), (4.11) with $D_N(\omega_k) = 0$ and the algorithm of equation (4.16) we can compute the bounding function $\bar{E}_{\text{cumf}, N}^n(\omega_k)$. In addition, for comparison, we compute the magnitude of the actual cumulative frequency-domain error function $|E_{\text{cumf}, N}^n(\omega_k)| = |G_{\text{cumf}, N}^n(\omega_k) - G_{\text{true}}(e^{j\omega_k T})|$. These frequency-domain functions are shown for two different cases of input signal, both of which satisfy $u_{\max} = 1$. In Fig. 3, we show the result of using a unity amplitude sine-wave, $\sin(2\pi n/5)$ for $n \geq 0$, at time index 111. Clearly, at the sine-wave frequency of 2 rad sec^{-1} the actual error and the bound are both reduced to near zero. Note that the initial transient that the plant experiences yields useful information at many frequencies other than just that of the sine wave. For frequencies lower than 2 rad sec^{-1} , the error bound is relatively tight. The opposite peaks at 4 rad sec^{-1} are due to the fact that in equation (4.6) the phase cancellations that occur in the

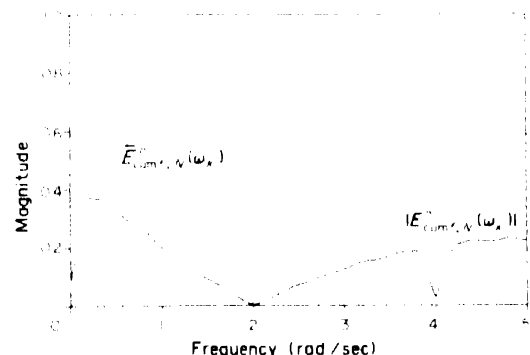


FIG. 3. Cumulative error $|E_{\text{cumf}, N}^n(\omega_k)|$ and bound $\bar{E}_{\text{cumf}, N}^n(\omega_k)$ for $n = 111$ and sinusoidal input.

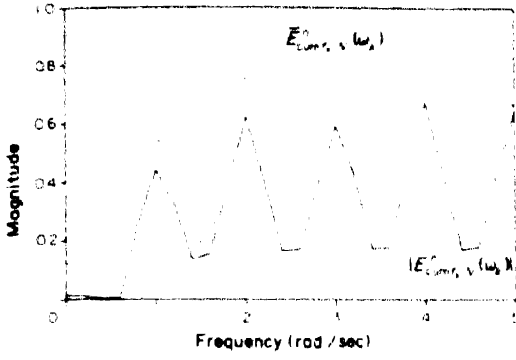


FIG. 4. Cumulative error $|E_{cumf, N}^n(\omega_k)|$ and bound $E_{cumf, N}^n(\omega_k)$ for $n = 111$ and square-wave input

actual frequency-domain error function of equation (2.4) are ignored and instead all terms are summed. In Fig. 4, we show the actual error and the bound using a unity-magnitude square-wave input with a fundamental frequency of 0.5 rad sec^{-1} . We see that equation (4.6) yields a tight bound since for this square-wave case the summation terms in equation (2.4) are nearly in phase. (In this example, the last term in equation (4.6) was only 0.078.) As expected, the frequency ranges where good identification occurs correspond to the frequency ranges where the input signal has its energy, namely the square-wave's fundamental frequency and its harmonics.

8.2 Closed-loop example description.

In this subsection, we will describe the components of the closed-loop adaptive control system of Fig. 2. These components are described so that we can examine the closed-loop performance of the robust estimator.

Plant

Consider the continuous-time *nominal* plant model given by the transfer function

$$G^r(s) = \frac{\omega_n^2 \left(\frac{s}{2\zeta\omega_n} + 1 \right)}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (8.3)$$

For $\zeta = 0.2$ and $\omega_n = 2 \text{ rad sec}^{-1}$ and adding some second-order unmodeled dynamics, we obtain

$$G_{true}^c(s) = \frac{4 \left(\frac{s}{0.8} + 1 \right)}{s^2 + 0.8s + 4} \cdot \frac{2500}{s^2 + 30s + 2500} \quad (8.4)$$

for the continuous-time *true* plant. Computing the zero-order hold equivalent of the *nominal* and *true* plants for a sampling period of $\pi/25 \text{ sec}$ yields

$$G(z) = \frac{0.62189z - 0.56211}{z^2 - 1.84458z + 0.90436} \quad (8.5)$$

and

$$G_{true}(z) = \frac{0.57666z^4 - 0.65109z^3 + 0.12783z^2}{z^4 - 2.13562z^3 + 1.46426z^2 - 0.30573z + 0.020849} \quad (8.6)$$

respectively.

In our simulations, the continuous-time parameter space is formed by letting ζ vary between 0.2 and 0.8 and ω_n vary between 1 and 2 rad sec^{-1} subject to the constraint that $\zeta\omega_n \geq 0.4 \text{ sec}^{-1}$ in the continuous-time nominal model. Then, to generate the discrete-time parameter space Θ of Assumption A1.2, a dense grid in ζ and ω_n is formed and the zero-order hold equivalent of the nominal model is computed for each $\zeta - \omega_n$ pair. In the simulations, all of the intermediate parameter vectors for the discrete-time nominal model are rounded to the nearest point in the resulting grid of discrete-time parameters. It was found empirically, that for all $\theta \in \Theta$, the discrete-time impulse response of the true plant was bounded by $0.75 \cdot 0.95^n$, for $n \geq 1$.

Disturbance signal

The output disturbance $d[n]$ in Fig. 2 is generated by passing a pseudo-random signal through a low-pass filter and then clipping the filter output such that the magnitude of the resulting signal is bounded by 0.01. The pseudo-random signal is uncorrelated in time and has a Gaussian probability distribution with zero mean and a standard deviation of 0.0075. The low-pass filter is given by

$$\frac{0.052881z^2}{1.64799z + 0.70087} \quad (8.7)$$

In this example, the disturbance signal is bounded in order for us to be able to determine a bound on $\hat{D}_N(\omega_k)$. Note that the disturbance signal has most of its energy in the same low frequency range that the magnitude of the plant is large (i.e. $\omega < 2 \text{ rad sec}^{-1}$).

Reference signal

The reference $r[n]$ in Fig. 2 is given by

$$r[n] = \begin{cases} \frac{1}{10}, & \text{if } \sin\left(\frac{\pi}{2000}(n \bmod 1000)^2\right) \geq 0, \\ -\frac{1}{10}, & \text{if } \sin\left(\frac{\pi}{2000}(n \bmod 1000)^2\right) < 0, \end{cases} \quad (8.8)$$

for $n \geq 0$. This periodic reference signal is a rich square-wave like signal whose frequency increases from zero (i.e. a step) to half of the Nyquist frequency (25 rad sec^{-1}) during a cycle

of 1000 samples. The reference signal was chosen to be rich so as to demonstrate the ability of the frequency-domain estimator to accurately identify a partially unknown plant and thus provide useful information to a control-law redesign algorithm. The use of a less rich, more bandlimited reference signal will result in good on-line identification only in the frequency range in which the input signal to the plant has significant energy. This might or might not lead to improved closed-loop performance depending upon the relative location of the frequency range where the *a priori* uncertainty is large.

Control-law redesign algorithm

In Fig. 2 we use the compensator $K(z)$ given by

$$K(z, \hat{\theta}) = G^{-1}(z, \hat{\theta}) \cdot \frac{c}{z-1}$$

where

$$c = \begin{cases} \frac{1}{2} \left(\frac{-1}{x} + \sqrt{\frac{1}{x^2} + \frac{4}{x}} \right), & \text{if } x > 0 \\ 1, & x \leq 0 \end{cases} \quad (8.9)$$

and

$$x = \sup_{\omega \in (0, \pi/T]} \left\{ \frac{[\Delta_{su}(e^{j\omega T}, \hat{\theta})^2 - 1]}{2(1 - \cos(\omega T))} \right\}$$

where $\hat{\theta}$ is the estimate yielded by the frequency-domain parameter estimator of Section 5. The above control-law attempts to cancel out the nominal plant dynamics and yield a discrete-time integrator for the loop transfer function. In equation (8.9), the choice of the gain constant c of the compensator is based on the continuous uncertainty bounding function $\Delta_{su}(e^{j\omega T}, \hat{\theta})$. A value for x , in equation (8.9), can be determined using the discrete bounding function $\Delta_{su}^n(e^{j\omega_k T}, \hat{\theta})$ and an appropriate additive safety factor as was discussed in Subsection 6.3.

The above control-law cannot be used for general plants since, as is shown in LaMaire (1987), it is only guaranteed to yield a robustly stable closed-loop system for nominal plants that have a relative degree of one or less. It is presented only so that the closed-loop performance of the robust estimator can be demonstrated. In the simulation examples that are described in Subsection 8.3 the compensator is updated every 100 samples or every 12.6 sec. That is, the current parameter estimates and uncertainty bounding function are used "infrequently" to design a new compensator.

Probing signal strategy

When the plant input signal is not rich, the robust estimator does not improve its estimates

and consequently, the control-law is not updated. If we want to enhance identification, that is, enable our robust estimator to reduce the frequency-domain uncertainty, we can add a probing signal in the closed-loop system (see Fig. 2).

In Fig. 2, the probing signal is generated by a weighted sum of sinusoids with randomly chosen phases φ_k .

$$v[n] = \frac{1}{N} \sum_{k=1}^{N/4} |V_N(\omega_k)| \cos \left(\frac{2\pi k n}{N} + \varphi_k \right),$$

for $n = 0, \dots, N-1$. (8.10)

We choose not to excite the system at high frequencies since we will then be exciting the unmodeled dynamics of the plant. We also do not excite the plant at $\omega = 0$ since we already know the D.C. gain of the true plant in our example. The following weights $|V_N(\omega_k)|$ for equation (8.10) are used in the simulation of Subsection 8.3

$$|V_N(\omega_k)| = \gamma \cdot \left(1 + \frac{c}{|e^{j\omega_k T} - 1|} [1 + \Delta_{su}^n(e^{j\omega_k T}, \hat{\theta})] \right) \frac{\bar{E}_{rem} + \bar{D}_N(\omega_k)}{e^{j\omega_k T} - 1 + 1 \inf_{\theta \in \Theta} \{|G(e^{j\omega_k T}, \theta)|\}} \quad (8.11)$$

where we set the design parameter γ to 10, c is the current compensator gain constant from equation (8.9), $c_r = 0.47$ is a design parameter that is chosen to be the target value for the compensator gain, and $\bar{E}_{rem} = 0.29$ is the last term (the remainder summation term) in equation 4.4. The second multiplicand in equation (8.11) accounts for the fact that the probing signal in Fig. 2 passes through the transfer function $[1 + G_{true}(z)K(z, \hat{\theta})]^{-1}$ before it reaches the input of the plant. Thus, the probing signal must be appropriately pre-emphasized to overcome this rejection by the closed-loop system. The third multiplicand in equation (8.11) gives the approximate shape of the magnitude of the desired input signal DFT $|U_N(\omega_k)|$. This term was developed using equations 4.4, 4.11 and 6.6 (assuming $U_N^{n-1}(\omega_k) \approx U_N^n(\omega_k)$ and $G_{cumf,N}^n(\omega_k) \approx G(e^{j\omega_k T}, \hat{\theta})$) and by noting that the following equality must be satisfied in order for the compensator gain c to achieve its target value c_r :

$$\Delta_{su}^n(e^{j\omega_k T}, \hat{\theta}) < \frac{e^{j\omega_k T} - 1}{c_r} + 1, \quad \forall \omega_k. \quad (8.12)$$

While the development of this probing signal algorithm is admittedly *ad hoc*, it does illustrate the possibility of using the evolving knowledge

of the uncertainty in the system to synthesize a tailored probing signal to reduce the remaining uncertainty.

8.3. Closed-loop simulations

In this subsection, we present the simulation results for the closed-loop adaptive control system of Fig. 2 that was described in Subsection 8.2. In our simulations, we use a DFT length N of 1000 points and a value of 200 for M in equation (4.4). [The fact that our choice of DFT length corresponds to the periodicity of the reference signal aids in the identification procedure since it helps reduce the differences in the shifted DFTs of the plant input signal in equation (4.4).] In addition, in Fig. 2, we assume that at the plant input there is a saturating actuator which constrains the input signal to lie between -1 and $+1$ (i.e. $u_{\max} = 1$).

We examine the performance of the adaptive control system of Fig. 2 for two cases, one without and one with the probing signal. For initial values, in both simulations, we set the cumulative frequency-domain estimate to the frequency response of the nominal model for $\zeta = 0.8$ and $\omega_n = 1 \text{ rad sec}^{-1}$, and set the corresponding cumulative frequency-domain error bounding function to the best bounding function that can be found using only *a priori* knowledge of the plant. Thus, we start the frequency-domain bounding method with parameter values that are very far from the true values of $\zeta = 0.2$ and $\omega_n = 2 \text{ rad sec}^{-1}$.

No probing signal case

In this simulation, we rely solely on the reference signal $r[n]$ for excitation of the plant. In Fig. 5, we show at time index 2500 (314.2 sec) the cumulative error bounding function $\bar{E}_{\text{cumf},N}^n(\omega_k)$ and, for comparison, the magnitude of the actual cumulative frequency-domain error function $|E_{\text{cumf},N}^n(\omega_k)| = |G_{\text{cumf},N}^n(\omega_k) - G_{\text{true}}(e^{j\omega_k T})|$. The effect of the disturbance is

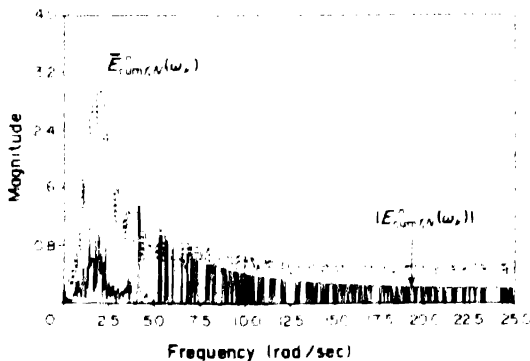


FIG. 5. Cumulative error $|E_{\text{cumf},N}^n(\omega_k)|$ and bound $\bar{E}_{\text{cumf},N}^n(\omega_k)$ for $n = 2500$, no probing signal case.

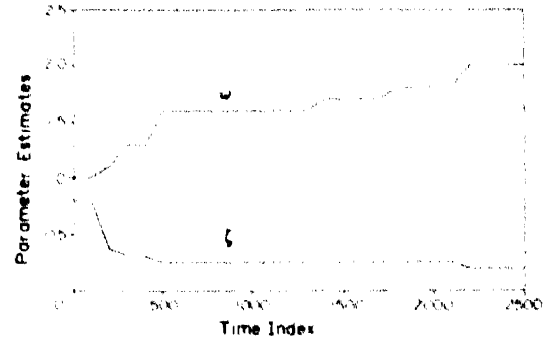


FIG. 6. Time histories of parameter estimates for no probing signal case

evident in the frequency range from 0 to 2.5 rad sec^{-1} . At higher frequencies, we can still see visages of the smooth *a priori* bounding function that was used to initialize the frequency-domain bounding method. In Fig. 6, we show the continuous-time parameters that correspond to the nearest point in the grid of discrete-time parameters that was described in Subsection 8.2. This continuous-time parameter interpretation allows us to see how well the frequency-domain estimator is identifying the plant as a function of time. Starting from the initial values of $\zeta = 0.8$ and $\omega_n = 1 \text{ rad sec}^{-1}$, which results from our *a priori* choice of the cumulative frequency-domain estimate, the parameters eventually reach their correct values of $\zeta = 0.2$ and $\omega_n = 2 \text{ rad sec}^{-1}$ at time index 2200 (276.5 sec). While the frequency-domain bounding methodology is continually in operation, the frequency-domain parameter estimator of Section 5 is only used at the control-law redesign times which occur every 100 samples (12.6 sec). This infrequent parameter updating accounts for the movement of the parameter estimates in jumps in Fig. 6. In order to see how well the robust estimator is performing as part of the adaptive control system, in Fig. 7 we show for the closed-loop system the nominal continuous-time bandwidth. We define this bandwidth

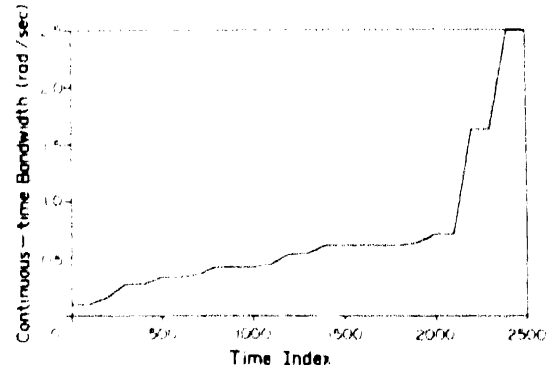


FIG. 7. Time history of continuous-time bandwidth for no probing signal case.

measure to be $(-1/T)\ln(1-c)$ where T is the sampling period of $\pi/25$ sec and c is the gain constant of the compensator $K(z)$ that was described previously. Note that if the inverted plant model in equation (8.9) were to exactly cancel the true plant, then the closed-loop transfer function of the system of Fig. 2 would be $c/(z-1+c)$, which has a stable pole at $1-c$ for $0 < c < 1$. Thus, Fig. 7 is simply a time history of the continuous-time pole corresponding to the discrete-time pole of the closed-loop nominal system. Comparing Figs 6 and 7, we see that as the robust estimator reduces the structured plant uncertainty, the compensator gain can be increased resulting in a nominal closed-loop bandwidth of about 2.5 rad sec^{-1} at time index 2500 (314.2 sec). Most of this improvement occurs rather late in the simulation after the second 1000 point cycle of the reference signal. To speed up this process in the next simulation we add the probing signal.

Probing signal case

In Figs 8–10 we show the analogous figures to those of the previous example, for the case of using both the reference signal and the probing signal algorithm that were described in Subsection 8.2. In this simulation the probing signal weights $V_N(\omega_k)$ were computed at time 0 and then recomputed at time index 1000 (125.7 sec) based on the current information. At time index 1100 (138.2 sec) the probing signal was disabled when the nominal closed-loop bandwidth exceeded our target value of 5 rad sec^{-1} . Comparing Figs 5 and 8 and noting the difference in scales, we see how the presence of the probing signal, which is concentrated between 0 and $12.5 \text{ rad sec}^{-1}$, has greatly reduced the frequency-domain estimation error. From Figs 6 and 9, we see that the robust estimator can identify the parameters much more quickly in the probing signal case than when the probing

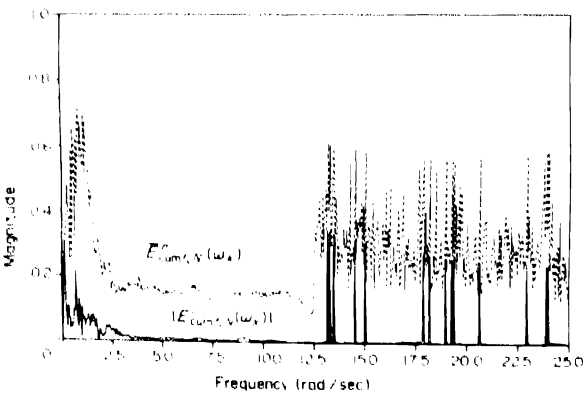


FIG. 8. Cumulative error $|E_{cum,y}^n(\omega_k)|$ and bound $E_{cum,y}^n(\omega_k)$ for $n = 2500$, probing signal case.

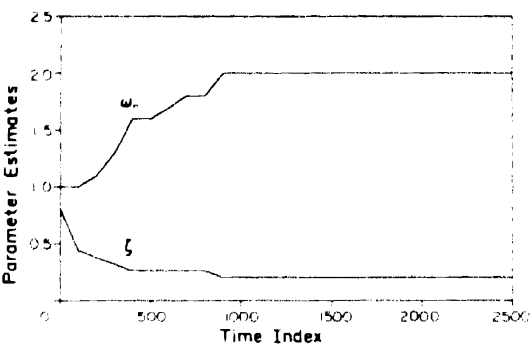


FIG. 9. Time histories of parameter estimates for probing signal case.

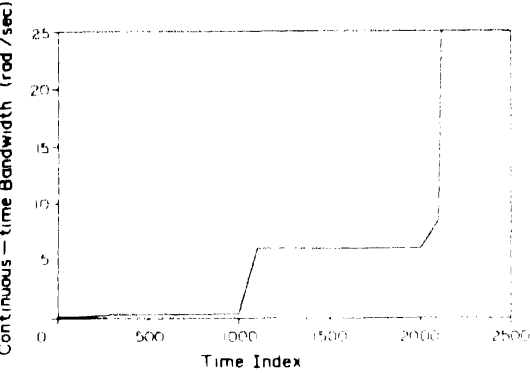


FIG. 10. Time history of continuous-time bandwidth for probing signal case.

signal was not present. In addition, comparing Figs 7 and 10, this improvement in identification speed is also reflected in an improved nominal closed-loop bandwidth. Note that at time index 1100 (138.2 sec) a nominal closed-loop bandwidth of about 6 rad sec^{-1} is achieved. After the probing signal is disabled, the system goes on to further reduce the frequency-domain uncertainty resulting in an increase of the compensator gain c to almost one (i.e. a deadbeat controller for the nominal plant). While the probing signal has improved the performance of the identification process and ultimately the closed-loop system, its presence also significantly degrades the command-following performance of the closed-loop system.

Comments

Several other simulation examples that had different values of nominal plant parameters were studied in order to more fully understand the performance of the robust estimator. The detailed results of these simulations are presented in LaMaire (1987). The primary conclusion drawn from our simulations was that a robust adaptive control system that uses the robust estimator can increase the closed-loop bandwidth and hence, improve the performance

of a system, under the right excitation conditions. Since a range of cases were considered, several different types of behavior were observed. In some situations, the reference signal provided sufficient excitation for the robust estimator to identify the plant well, resulting in the achievement of the target closed-loop bandwidth. However, in other identification cases (i.e. in cases where the initial compensator was chosen such that frequencies in the range of the target closed-loop bandwidth were greatly attenuated resulting in little excitation of the plant at these frequencies), the reference signal itself had to be supplemented by the aforementioned probing signal $v[n]$ in order for the robust estimator to be able to identify the plant well enough to increase the closed-loop bandwidth. Thus, in a closed-loop context, it was our experience that given excitation at the proper frequencies, the robust estimator was able to yield an improved nominal plant model and uncertainty bound so that the robust control-law redesign algorithm could increase the bandwidth of the closed-loop system.

9. CONCLUDING REMARKS

In this paper, we presented a new estimation (identification) methodology that can be used in a robust adaptive controller to provide stability-robustness guarantees. The key feature of the robust estimator is the frequency-domain bounding function on the modeling uncertainty. Our simulation results revealed that the use of the robust estimator can yield improved closed-loop performance, that is, increased bandwidth as compared with the best LTI compensator that could have been designed (for a given design method) using only *a priori* knowledge of the plant. In some situations, the plant input signal is not rich enough to allow identification. In these cases, one can choose to either use the best *a priori* control-law or introduce an external probing signal to enhance identification. As a final remark, we note that while the robust estimator provides guarantees that no other methodology can, the price of these guarantees is the large computational load of the frequency-domain calculations described in Section 4.

Acknowledgements—The authors are grateful to F. C. Schweppe, L. Ljung and an anonymous reviewer for their insightful comments. This study was supported by the NASA Ames and Langley Research Centers under grant NASA/NAG-2-297, by the Office of Naval Research under contract ONR/N00014-82-K-0582 (NR 606-003) and by the National Science Foundation under grant NSF/ECS-8210960.

REFERENCES

- Athans, M. (1986). A tutorial on the LQG/LTR method. *Proc. Amer. Control Conf.*, Seattle, WA, pp. 1289–1296.
- Goodwin, G. C., D. J. Hill and M. Palamizwami (1985a). Towards an adaptive robust controller. *Proc. IFAC Identification and System Parameter Estimation Conf.*, York, U.K., pp. 997–1002.
- Goodwin, G. C., D. J. Hill, D. O. Mayne and R. H. Middleton (1985b). Adaptive robust control (convergence, stability and performance). Technical Report EF8544, Univ. of Newcastle, N.S.W. 2308, Australia.
- Goodwin, G. C., R. Lozano-Leal, D. O. Mayne and R. H. Middleton (1986). Rapprochement between continuous and discrete model reference adaptive control. *Automatica*, **22**, 199–207.
- Kosut, R. L. (1986). Adaptive calibration: An approach to uncertainty modeling and on-line robust control design. *Proc. 25th Conf. on Decision and Control*, pp. 455–461.
- Kosut, R. L. (1987). Adaptive uncertainty modeling, on-line robust control design. *Proc. Amer. Control Conf.*, pp. 245–250.
- Kosut, R. L. (1988). Adaptive control via parameter set estimation. *Int. J. Adaptive Control Signal Process.*, **2**, 371–399.
- LaMaire, R. O. (1987). Robust time and frequency domain estimation methods in adaptive control. Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA, U.S.A.
- LaMaire, R. O., I. Valavani, M. Athans and G. Stein (1987). A frequency-domain estimator for use in adaptive control systems. *Proc. Amer. Control Conf.*, pp. 238–244.
- Lehtomäki, N. A., D. A. Castanon, B. C. Levy, G. Stein, N. R. Sandell, Jr and M. Athans (1984). Robustness and modeling error characterization. *IEEE Trans. Aut. Control*, **AC-29**, 212–220.
- Ljung, L. (1985). On the estimation of transfer functions. *Automatica*, **21**, 677–698.
- Ljung, L. (1987). *System Identification—Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ.
- Morrison, S. and B. Walker (1988). Batch-least-squares adaptive control in the presence of unmodelled dynamics. *Proc. Amer. Control Conf.*, pp. 774–776.
- Oppenheim, A., A. Willsky and I. Young (1983). *Signals and Systems*. Prentice-Hall, Englewood Cliffs, NJ.
- Parker, P. J. and R. R. Bitmead (1987a). Approximation of stable and unstable systems via frequency response interpolation. *Proc. IFAC 10th Triennial World Congr.*, pp. 357–362.
- Parker, P. J. and R. R. Bitmead (1987b). Adaptive frequency response identification. *Proc. 26th Conf. on Decision and Control*, pp. 348–353.
- Strang, G. (1980). *Linear Algebra and its Applications*. Academic Press, New York.

APPENDIX. DERIVATION OF SIGNAL PROCESSING THEOREMS

In this appendix, we provide proofs for Theorems 1 and 2. Further, we state and prove two additional theorems, the last of which can be used to compute a time-domain bound on the output of a linear system by using the DFT of the input of the system. This result is referred to in Section 7 of the paper.

Proof of Theorem 1. We know that

$$Y(e^{j\omega_k T}) = H(e^{j\omega_k T})U(e^{j\omega_k T}), \quad \text{for } k = 0, \dots, N-1, \quad (\text{A.1})$$

where $U(e^{j\omega_k T})$ and $Y(e^{j\omega_k T})$ and the DTFTs of $u[n]$ and $y[n]$, respectively. Since

$$Y(e^{j\omega_k T}) = \sum_{m=-\infty}^{\infty} y[m]W_N^{km} + Y_N^* \omega_k + \sum_{m=-\infty}^{\infty} y[m]W_N^{km}, \quad \text{for } k = 0, \dots, N-1, \quad (\text{A.2})$$

and a similar expression holds for $U(e^{j\omega_k T})$, we can write

$$Y_N^n(\omega_k) = H(e^{j\omega_k T}) \times \left\{ \sum_{m=n-N}^n u[m] W_N^{km} + U_N^n(\omega_k) + \sum_{m=n+1}^n u[m] W_N^{km} \right\} - \left\{ \sum_{m=n-N}^n y[m] W_N^{km} + \sum_{m=n+1}^n y[m] W_N^{km} \right\},$$

for $k = 0, \dots, N-1$. (A.3)

It can be shown that

$$\sum_{m=n-N}^n y[m] W_N^{km} = h[0] \left\{ \sum_{m=n-N}^n u[m] W_N^{km} \right\} + \sum_{p=1}^N h[p] W_N^{kp} \left\{ \sum_{m=n-N}^n u[m] W_N^{km} - \sum_{m=n-N}^n u[m] W_N^{km} \right\},$$

for $k = 0, \dots, N-1$. (A.4)

So,

$$H(e^{j\omega_k T}) \sum_{m=n-N}^n u[m] W_N^{km} - \sum_{m=n-N}^n y[m] W_N^{km} = \sum_{p=1}^N h[p] W_N^{kp} \left\{ \sum_{m=n-N}^n u[m] W_N^{km} \right\},$$

for $k = 0, \dots, N-1$. (A.5)

Similarly,

$$H(e^{j\omega_k T}) \sum_{m=n+1}^n u[m] W_N^{km} - \sum_{m=n+1}^n y[m] W_N^{km} = - \sum_{p=1}^N h[p] W_N^{kp} \left\{ \sum_{m=n+1}^n u[m] W_N^{km} \right\},$$

for $k = 0, \dots, N-1$. (A.6)

Using equations (2.3), (A.3) and (A.5, A.6) we find that

$$E_N^n(\omega_k) = \sum_{p=1}^N h[p] W_N^{kp} \left\{ \sum_{m=n-N}^n u[m] W_N^{km} - \sum_{m=n-N}^n u[m] W_N^{km} \right\} = \sum_{p=1}^N h[p] W_N^{kp} \left\{ \sum_{m=n-N}^n u[m] W_N^{km} - \sum_{m=n-N}^n u[m] W_N^{km} \right\},$$

for $k = 0, \dots, N-1$. (A.7)

Equation (2.4) now follows using the definition of equation (2.1). Q.E.D.

Proof of Theorem 2 Using the triangle inequality and equations (2.4) and (A.7) we find

$$|E_N^n(\omega_k)| \leq \sum_{p=1}^{M-1} |h[p]| |U_N^n(\omega_k) - U_N^n(\omega_k)| + \sum_{p=M}^N |h[p]| \left| \sum_{m=n-N}^n u[m] W_N^{km} - \sum_{m=n-N}^n u[m] W_N^{km} \right|,$$

for $k = 0, \dots, N-1$. (A.8)

$$\text{Since } \left| \sum_{m=n-N}^n u[m] W_N^{km} - \sum_{m=n-N}^n u[m] W_N^{km} \right| \leq \sum_{m=n-N}^n |u[m]| + \sum_{m=n-N}^n |u[m]| \leq 2u_{\max} p \quad (\text{A.9})$$

we conclude that equation (2.5) is true. Q.E.D.

Corollary 1. Under the assumptions of Theorem 2,

$$|E_N^n(\omega_k)| \leq 2u_{\max} \sum_{p=1}^N p |h[p]|, \quad \text{for } k = 0, \dots, N-1. \quad (\text{A.10})$$

Proof. Choose $M = 1$ in Theorem 2. This corollary is closely related to Theorem 2.1 in Ljung (1987).

The following theorems are useful for computing the maximum output signal of a transfer function for which we have a magnitude bounding function in the frequency domain.

Theorem A.1. Let $y[n] = h[n] * u[n]$, where $h[n]$ is an infinite-length, causal, impulse response with all its poles in the open unit disk. We denote the DTFT of $h[n]$ by $H(e^{j\omega_k T})$, and the DFT of the N -points of $u[n]$ ending with time index n , by $U_N^n(\omega_k)$. Then

$$y[n] = \frac{1}{N} \sum_{k=0}^{N-1} H(e^{j\omega_k T}) U_N^n(\omega_k) W_N^{-kn} + e[n], \quad (\text{A.11})$$

where

$$e[n] = \sum_{p=N}^{\infty} h[p] (u[n-p] - u[n-(p \text{ modulo } N)]). \quad (\text{A.12})$$

Remark A.1. The signal $e[n]$ is the error due to the fact that the impulse response $h[n]$ is of infinite length. We note from equation (A.12) that if $h[p] = 0$ for $p \geq N$, then $e[n] = 0$, $\forall n$.

Proof. See LaMaire (1987).

We want to be able to find a magnitude bounding function on $y[n]$. The following theorem provides such a bounding function by using the results of Theorem A.1.

Theorem A.2. Under the assumptions of Theorem A.1 we find that, for a real-valued impulse response $h[n]$ and a real-valued signal $u[n]$, the magnitude of $y[n]$ is bounded at each n as follows:

$$|y[n]| \leq \frac{1}{N} \left\{ |H(e^{j\omega_0 T})| |U_N^n(\omega_0)| + 2 \sum_{k=1}^{(N/2)-1} |H(e^{j\omega_k T})| |U_N^n(\omega_k)| + |H(e^{j\omega_{(N/2)} T})| |U_N^n(\omega_{(N/2)})| \right\} + 2u_{\max} \sum_{p=N}^{\infty} |h[p]|, \quad (\text{A.13})$$

where

$$u_{\max} = \sup_m |u[m]|, \quad (\text{A.14})$$

and where we have assumed that N is even. An alternate form of the theorem can easily be proven for the case of an odd value of N .

Proof. See LaMaire (1987).

Enhancement of Fixed Controllers via Adaptive- Q Disturbance Estimate Feedback*

T. T. TAY† and J. B. MOORE‡

Robust and adaptive control techniques are blended to enhance one another in what is termed an adaptive- Q scheme.

Key Words—Adaptive control, adaptive systems, disturbance rejection, least squares estimation, robust control, optimal control.

Abstract—A direct adaptive control scheme is proposed, termed an Adaptive- Q scheme, which has the property that it can, if considered desirable, limit its searches to the space of all stabilizing linear proper controllers for a nominal linear plant. The adaptive controller feeds back disturbance estimates via a filter with operator Q . It augments any fixed linear stabilizing controller, viewed as a minimum variance disturbance decoupling controller, for the nominal plant. For a nominal plant, the adaptive- Q augmentations evanesce, but when the plant is other than the nominal one, the adaptive- Q scheme seeks to minimize the effect of disturbances and thus to improve the performance over that of the fixed controller.

The proposed adaptive- Q disturbance estimate feedback (DEF) controllers can be simple to implement even for high order multivariable plants with high order fixed controllers, and have the significance that they seek to enhance performance of standard controller designs in the face of plant perturbations or uncertainties, rather than supplant or compete with them. Prior knowledge about the frequency band of plant uncertainty is readily incorporated into the adaptive design scheme.

1. INTRODUCTION

ADAPTIVE CONTROLLERS for simple processes are well studied, and indeed under certain idealized conditions are known to be globally convergent (Goodwin *et al.*, 1980; Narendra *et al.*, 1980; Moore, 1985). When applied in many situations which are less than ideal, they have at best local convergence properties (Kosut and Anderson, 1984). In current adaptive control research, there is a move away from seeking “universal” schemes for application to any control situation to more limited objectives within restricted

classes of control applications. An area which offers scope for achieving significant practical results is the blending of adaptive techniques with robust or optimal fixed controller design (see for example Moore *et al.*, 1982). The idea is to exploit the strength of fixed off-line robust/optimal controller design for complex systems and the capability of adaptive techniques to give on-line performance enhancement in the face of plant perturbations or uncertainties.

In earlier studies (Chakravarty and Moore, 1986), one of the authors demonstrated that simple adaptive pole assignment or linear quadratic Gaussian schemes could be applied to enhance the stability and performance properties of a high order multivariable robust/optimal (frequency shaped linear quadratic Gaussian) off-line controller design. The blending of the on-line and off-line techniques is based on engineering insights of the particular problem. The success of the blending technique in Chakravarty and Moore (1986) depends on the fact that the plant uncertainties are restricted to a narrow frequency band and have a low order representation. The question arises as to the potential of blending off-line and on-line control techniques in more general situations. One objective in a blending procedure is for the on-line and off-line designs to work towards either the same performance objective, or to complementary objectives. In the former case, should the plant be the nominal one, the adaptive scheme should be tuned so that it does not provide any additional control asymptotically. An example of complementary objectives is that the off-line design be robust and the on-line design enhance performance, even for the nominal plant case.

In this paper, we propose a novel class of adaptive disturbance estimate feedback (DEF) controller. The objective is to blend on-line

* Received 14 July 1987; revised 15 March 1988; revised 30 November 1988; received in final form 4 May 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor L. Valavani under the direction of Editor P. C. Parks.

† Department of Electrical Engineering, National University of Singapore, Kent Ridge 0521, Singapore.

‡ Department of Systems Engineering, Research School of Physical Sciences, Australian National University, P.O. Box 4, Canberra A.C.T. 2600, Australia. Author to whom all correspondence should be addressed.

adaptive control techniques with off-line designed controllers so as to enhance the performance of the latter. The understandable fear of any fixed robust controller designer is that in superimposing adaptive loops, the adaptations based on crude calculations could dominate and destroy the control action of the carefully designed off-line controller. To limit the adaptive control actions when indicated, we impose the following requirements for the adaptive loop in the first instance.

(i) When the plant is the nominal one, there is no adaptive control action, at least asymptotically.

(ii) The adaptive scheme seeks to minimize the same disturbance rejection criterion that the off-line design minimizes for the nominal plant.

(ii) The adaptive scheme searches only over the class of stabilizing proper controllers for the nominal plant, when so instructed.

In Section 2, any off-line designed stabilizing proper linear controller for a nominal linear plant is viewed as a minimum variance controller for some disturbance rejection index. As well, the class of all stabilizing proper controllers for the nominal plant is defined in terms of this perturbed stabilizing controller. The results in this section are developed for the multivariable case since they do not significantly increase the complexity of the presentation. In Section 3, the objectives of an on-line adaptive scheme for enhancing on-line performance are set out, and specific proposals are made for implementation. Here the results are presented for the scalar case only, noting that the ideas have potential for dealing with the multivariable case. Certain basic properties of the proposed blending of on-line and off-line controller design are studied in Section 4 with more properties studied in subsequent work. Simulation examples are studied in Section 5, including ones which make connections to earlier work. Conclusions are drawn in Section 6.

2. CONTROLLER THEORY

In this section, background controller theory is organized to set up the motivation for the adaptive schemes of subsequent sections. We consider in turn plant descriptions, stabilizing controllers, the class of all stabilizing proper controllers, performance indices, inverse problems and disturbance rejection controllers for perturbed plants.

Plant description

Consider a plant with nominal state space description

$$x_{k+1} = Ax_k + Bu_k, \quad y_k = Cx_k + Du_k \quad (2.1)$$

and transfer function

$$G_0(z) = C(zI - A)^{-1}B + D = \begin{bmatrix} A & B \\ C & D \end{bmatrix}_T \in R_p \quad (2.2)$$

where R_p denotes the class of rational proper transfer functions, and $[\]_T$ denotes a transfer function as defined in (2.2) using block partitioning. Consider also coprime factorizations

$$G_0 = N_0 M_0^{-1} = \tilde{M}_0^{-1} \tilde{N}_0; \quad N_0, M_0, \tilde{N}_0, \tilde{M}_0 \in RH^\infty \quad (2.3)$$

where RH^∞ denotes the class of all asymptotically stable rational proper transfer functions. Suitable selections are given below in (2.7).

Stabilizing proper controllers

Consider proper stabilizing controllers for G_0 as $K \in R_p$ (rational proper transfer functions) [see Fig. 1], where the closed loop system is well posed in that $KG_0, G_0K \in R_{sp}$ (the class of strictly proper rational transfer functions). Thus

$$\begin{bmatrix} I & -K \\ -G_0 & I \end{bmatrix}^{-1} \text{ exists and belongs to } RH^\infty. \quad (2.4)$$

Consider also coprime factorizations for some stabilizing controller K_0 as

$$K_0 = U_0 V_0^{-1} = \tilde{V}_0^{-1} \tilde{U}_0; \quad U_0, V_0, \tilde{U}_0, \tilde{V}_0 \in RH^\infty \quad (2.5)$$

which satisfy the double Bezout equation

$$\begin{bmatrix} \tilde{V}_0 & -\tilde{U}_0 \\ -\tilde{N}_0 & \tilde{M}_0 \end{bmatrix} \begin{bmatrix} M_0 & U_0 \\ N_0 & V_0 \end{bmatrix} = \begin{bmatrix} M_0 & U_0 \\ N_0 & V_0 \end{bmatrix} \begin{bmatrix} \tilde{V}_0 & -\tilde{U}_0 \\ -\tilde{N}_0 & \tilde{M}_0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \quad (2.6)$$

Two factorizations

For some stabilizing constant state feedback gain F for (2.1) and some stabilizing constant output injections H, AL for (2.1), we consider two alternative factorizations satisfying (2.3)–

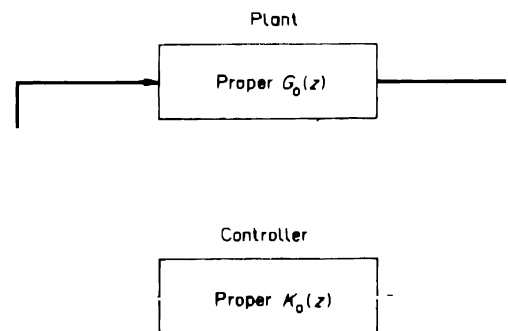


FIG. 1. Nominal plant and controller.

(2.6) as follows

$$\begin{bmatrix} M_0 & U_0 \\ N_0 & V_0 \end{bmatrix} = \begin{bmatrix} A+BF & B & -H \\ F & I & 0 \\ C+DF & D & I \end{bmatrix}_T, \\ \begin{bmatrix} A+BF & B & -(A+BF)L \\ F & I & -FL \\ C & 0 & I \end{bmatrix}_1 \in RH^* \\ \begin{bmatrix} \tilde{V}_0 & -\tilde{U}_0 \\ -\tilde{N}_0 & \tilde{M}_0 \end{bmatrix} = \begin{bmatrix} A+HC & -(B+HD) & H \\ F & I & 0 \\ C & -D & I \end{bmatrix}_1, \\ \begin{bmatrix} A+ALC & -B & AL \\ F+FLC & I & FL \\ C & 0 & I \end{bmatrix}_1 \in RH^*. \quad (2.7)$$

That the first factorizations in (2.7) satisfy (2.6) is known [see Moore *et al.* (1990) for generalizations and its references]. That the second factorizations satisfy (2.6) is shown in the Appendix with proofs following a similar pattern as that for the first factorizations. Notice that the first factorizations have an implicit requirement that $K_0 \in R_{sp}$ whereas the second requires that $G_0 \in R_{sp}(D=0)$.

Class of all stabilizing proper controllers

This class can be characterized (Moore *et al.*, 1990) in terms of an arbitrary $Q \in RH^*$ under (2.3), (2.5) and (2.6) as

$$K = UV^{-1}, \quad U = U_0 + M_0Q, \quad V = V_0 + N_0Q \\ = \tilde{V}^{-1}\tilde{U}, \quad \tilde{U} = \tilde{U}_0 + Q\tilde{M}_0, \quad \tilde{V} = \tilde{V}_0 + Q\tilde{N}_0. \quad (2.8)$$

Note also that (2.8) can be written via (2.6) as

$$K = K_0 + \tilde{V}_0^{-1}Q(I + V_0^{-1}N_0Q)^{-1}V_0^{-1} \quad (2.9)$$

which can be reorganized as in Fig. 2 with

$$J_0 = \begin{bmatrix} K_0 & \tilde{V}_0^{-1} \\ V_0^{-1} & -V_0^{-1}N_0 \end{bmatrix} \in R_p, \\ S_0 = \begin{bmatrix} I & -K_0 \\ -G_0 & I \end{bmatrix}^{-1} \begin{bmatrix} M_0 \\ N_0 \end{bmatrix} \in RH^*. \quad (2.10)$$

It is known that the associated four closed loop transfer functions between the u_i and e_i of Fig. 2a are affine in Q as follows

$$\begin{bmatrix} I & -K \\ -G_0 & I \end{bmatrix} = \begin{bmatrix} I & -K_0 \\ -G_0 & I \end{bmatrix}^{-1} + \begin{bmatrix} M_0 \\ N_0 \end{bmatrix} \\ \times Q[N_0 \quad \tilde{M}_0] \in RH^*. \quad (2.11)$$

(The inverses exist trivially when either G or $K_0 \in R_{sp}$.)

More generally the following result is readily established:

Lemma 1 Part (i). With $K_0 \in R_p$ stabilizing $G_0 \in R_p$ and J_0 defined in (2.10) the transfer function relating $u = [u_1' u_2' u_3' u_4']'$ to $e = [e_1' e_2' e_3' e_4']'$ of Fig. 2b is as follows

$$W = \begin{bmatrix} S_{11} & S_{12} & 0 \\ 0 & I & 0 \\ S_{21} & 0 & -I \end{bmatrix} + \begin{bmatrix} S_{12} \\ I \\ 0 \end{bmatrix} Q[S_{21} \quad 0 \quad I] \quad (2.12)$$

which is affine in Q . Moreover the system of Fig. 2 with G_0, J_0, Q having stabilizable and detectable realization is internally (asymptotically) stable if and only if Q is (asymptotically) stable. Further $W \in RH^*$ if and only if $Q \in RH^*$. **Part (ii).** Should G_0, J_0 be stabilizable and detectable realizations as in Part (i), but Q be some causal time-varying operator, then W generalizes as a causal time-varying operator.

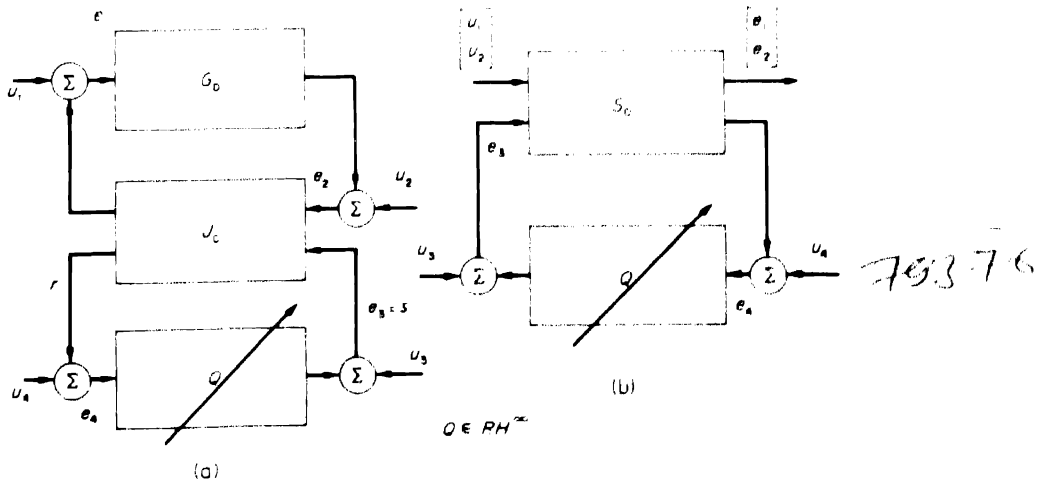


FIG. 2. Class of all stabilizing controllers.

Moreover, a necessary and sufficient condition for W to be a bounded-input, bounded-output operator is that Q be bounded-input, bounded-output.

Proof. See Appendix.

Remarks. This lemma is a natural generalization of earlier results for single loop time-invariant systems. It tells us that in implementing a control system as in Fig. 2, the operator Q can be any adaptive loop (for example), and as long as Q is causal bounded-input, bounded-output, then bounded-input, bounded-output stability of the system of Fig. 2 is maintained.

For the factorizations (2.7), it is immediate that

$$J_0 = \begin{bmatrix} A + BF + HC + HDF & -H & (B + HD) \\ F & 0 & I \\ -(C + DF) & I & -D \end{bmatrix}_T, \quad (2.13)$$

$$\begin{bmatrix} A + BF + ALC + BFLC & -(AL + BFL) & B \\ F + FLC & -FL & I \\ -C & I & 0 \end{bmatrix}_T$$

and the class of all stabilizing controllers can be organized as in Fig. 3.

Performance index

As a performance index associated with G , let us consider minimization of a disturbance response. Consider the plant G_0 augmented as $P \in R_p$ with $P_{22} = G_0$. Thus with Z-transform relationships, ignoring initial conditions, consider

$$\begin{bmatrix} e(z) \\ y(z) \end{bmatrix} = \begin{bmatrix} P_{11}(z) & P_{12}(z) \\ P_{21}(z) & P_{22}(z) \end{bmatrix} \begin{bmatrix} w(z) \\ u(z) \end{bmatrix} \quad (2.14)$$

$$P_{22} = G_0 \in R_p.$$

Here w_k is an unknown disturbance input, and e_k is a reference or disturbance response, frequently including terms involving u_k and x_k . [The case when w_k is a known (reference) input, or has such components is studied in a companion paper (Tay and Moore, 1990).]

Now with feedback controllers K applied to the plant G_0 so that $u(z) = K(z)y(z)$ then (2.14) can be reorganized (refer to Fig. 4) as

$$e(z) = F_K(z)w(z), \quad F_K = P_{11} + P_{12}K(I - G_0K)^{-1}P_{21}$$

$$e(z) = F_Q(z)w(z), \quad F_Q = T_{11} + T_{21}QT_{21} \text{ (affine in } Q\text{)} \quad (2.15)$$

where the second formulation follows from structuring K in terms of some K_0, J_0 and $Q \in RH^\infty$ as in (2.8)–(2.11), and re-organizing F_K of Fig. 4a as F_Q of Fig. 4b. This re-organizing

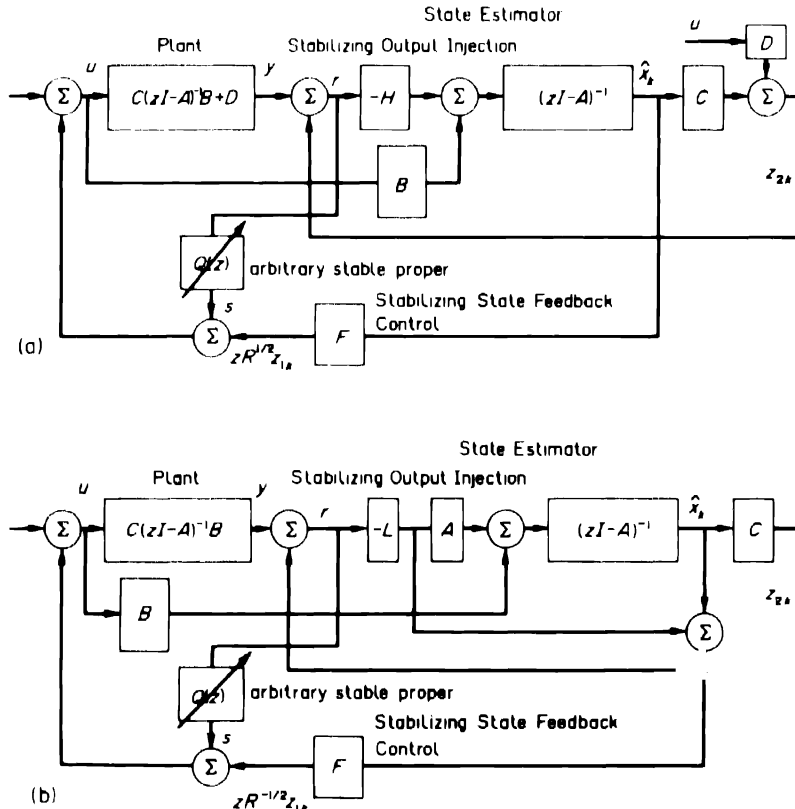


FIG. 3. Class of all stabilizing controllers based on state estimate feedback design.

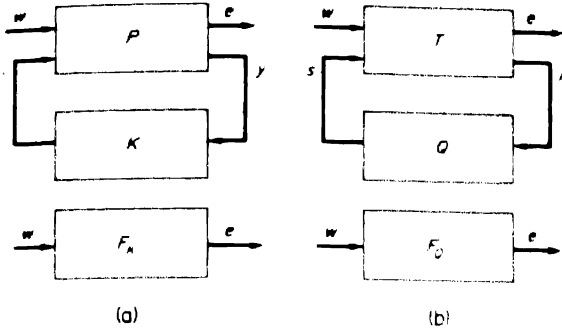


FIG. 4. Disturbance rejection controller.

leads to the following expression for T (note that $T_{22} = 0$).

$$T(G) = \begin{bmatrix} P_{11} + P_{12}K_0(I - G_0K_0)^{-1}P_{21} & P_{12}M_0 \\ \tilde{M}_0P_{21} & 0 \end{bmatrix} \quad (2.16)$$

Lemma 2.

$$K_0 \in R_p \text{ stabilizes } P \text{ (with } P_{22} = G_0) \quad (2.17)$$

if and only if

$$K_0 \in R_p \text{ stabilizes } G_0 \text{ and } T \in RH^+. \quad (2.18)$$

Moreover, under these conditions and with $Q \in RH^+$

$$F_{K_0} = T_{11} \in RH^+. \quad (2.19)$$

Proof. See Appendix.

Lemma 3. Consider the case of a plant G (not the nominal plant G_0) and in particular consider the scheme of Fig. 4 under (2.16) with the augmented plant $P(G) = P$ of (2.14) save that $P_{22} = G$ (true plant) instead of G_0 . Then the rearrangement is as in Fig. 5 with

$$T(G) = \begin{bmatrix} P_{11} + P_{12}K_0(I - GK_0)^{-1}P_{21} \\ V_0^{-1}(I - GK_0)^{-1}P_{21} \\ P_{12}(I - K_0G)^{-1}\tilde{V}_0^{-1} \\ V_0^{-1}(I - GK_0)^{-1}G\tilde{V}_0^{-1} - V_0^{-1}N_0 \end{bmatrix} \quad (2.20)$$

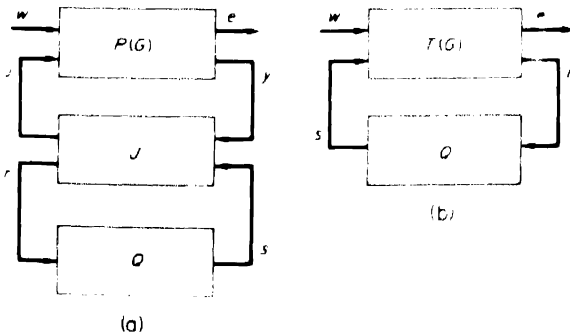


FIG. 5. Non-nominal plant case.

and

$$\begin{aligned} e(z) &= P_{11}w(z) + [P_{12}U_0 + P_{21}M_0Q]r(z), \\ r(z) &= \tilde{M}_0y(z) - \tilde{N}_0u(z) \\ &= P_{11}w(z) + P_{12}U_0r(z) + P_{12}M_0s(z), \\ s(z) &= Qr(z) \end{aligned} \quad (2.21)$$

Proof. See Appendix.

A minimum variance disturbance rejection task is as follows

$$\min_{\text{stabilizing } K} \| [F_K]_{\text{sp}} \|_2^2 = \min_{Q \in RH^+} \| [F_Q]_{\text{sp}} \|_2^2 \quad (2.22a)$$

where $[\cdot]_{\text{sp}}$ denotes the strictly proper part. Let us denote optimal K , Q as K^* , Q^* . When F_k and $F_Q \in R_{\text{sp}}$, the equivalent minimum variance task is as follows,

$$\min_{\text{stabilizing } K} E[e_k'e_k], \quad E[w_k] = 0, \quad E[w_k'w_k] = I. \quad (2.22b)$$

Example 1. Consider a stochastic version of the plant (2.1) in innovation representation form as

$$x_{k+1} = Ax_k + Bu_k + \Gamma w_k, \quad y_k = Cx_k + w_k.$$

Consider also a standard linear quadratic (LQG) index

$$I_{\text{LQG}} = E \left\{ \frac{1}{k} \sum_{i=1}^k (u_{i-1}'R_i u_{i-1} + x_i'Q_i x_i) \right\}, \quad R_i > 0, \quad Q_i \geq 0.$$

Now define

$$e_k = \begin{bmatrix} R_i^{1/2}u_{k-1} \\ Q_i^{1/2}x_k \end{bmatrix}, \quad P = \begin{bmatrix} 0 & 0 & 0 & R_i^{1/2} \\ 0 & A & \Gamma & B \\ I & 0 & 0 & 0 \\ 0 & Q_i^{1/2} & 0 & 0 \\ 0 & C & I & 0 \end{bmatrix}_T.$$

With the factorizations of (2.7b), then $T_{22} = 0$, $T_{21} = I$ and

$$T = \left[\begin{array}{cc|cc} 0 & R_i^{1/2}F & -R_i^{1/2}FL & R_i^{1/2} \\ 0 & A + BF & -(A + BF)L & B \\ \hline I & 0 & 0 & 0 \\ 0 & Q_i^{1/2} & 0 & 0 \\ 0 & 0 & I & 0 \end{array} \right]_T.$$

With the above definitions, $I_{\text{LQG}} = E\{k^{-1}\Sigma_k^t e_i'e_i\}$ and it is straightforward to see that the above LQG controller design task is equivalent to the disturbance rejection optimization (2.22) under (2.15).

Example 2. Consider the following plant description based on ARMAX (input/output)

scalar variable models

$$A(z)y(z) = B(z)u(z) + C(z)w(z)$$

where $A(z) = 1 + a_1 z^{-1} + \dots + a_n z^{-n}$, $B(z) = b_1 z^{-1} + \dots + b_m z^{-m}$, $C(z) = 1 + c_1 z^{-1} + \dots + c_1 z^{-n}$ and $G_0 = B(z)/A(z) \in R_{sp}$

$$e_k = \begin{bmatrix} R_c^{1/2} u_{k-1} \\ y_k \end{bmatrix}, \quad P = \begin{bmatrix} \begin{bmatrix} 0 \\ \frac{C(z)}{A(z)} \end{bmatrix} & \begin{bmatrix} z^{-1} R_c^{1/2} \\ \frac{B(z)}{A(z)} \end{bmatrix} \\ \begin{bmatrix} \frac{C(z)}{A(z)} \\ \frac{B(z)}{A(z)} \end{bmatrix} & \begin{bmatrix} z^{-1} R_c^{1/2} \\ \frac{B(z)}{A(z)} \end{bmatrix} \end{bmatrix}.$$

An optimal one-horizon nominal controller K_0 for (2.16) designed according to the LQ index (one horizon)

$$I_{1,0} = E\{y_{k+1}^2 + R_c u_k^2\} \quad R_c > 0$$

is given as follows [10]

$$K_0(z) = \frac{[A(z) - C(z)]z}{zB(z) + b_1^{-1}R_c(z)}.$$

Suitable factorizations for G_0 and K_0 are then given as follows

$$\begin{aligned} N_0 &= B(z)/C(z), \quad M_0 = A(z)/C(z) \\ U_0 &= \frac{[A(z) - C(z)]z}{zB(z) + b_1^{-1}R_c A(z)} \\ V_0 &= \frac{zB(z) + b_1^{-1}R_c C(z)}{zB(z) + b_1^{-1}R_c A(z)}. \end{aligned}$$

With these factorizations, then $T_{22} = 0$, $T_{21} = I$ and

$$T = \begin{bmatrix} \begin{bmatrix} z^{-1} R_c^{1/2} U_0 \\ V_0 \end{bmatrix} & \begin{bmatrix} z^{-1} R_c^{1/2} M_0 \\ N_0 \end{bmatrix} \\ 1 & 0 \end{bmatrix}.$$

Remark. It is noted that the selection of $P(G)$ to realize a particular disturbance response for a particular plant G and controller K is not unique. We will present subsequently other selections of $P(G)$.

The inverse problem of optimal disturbance rejection

It is known that not all stabilizing controllers are optimal in an LQ sense, as above, but for the more general index (2.22), we claim the following.

Lemma 4. Any stabilizing controller $K_0 \in R_p$ for $G_0 \in R_{sp}$ is optimal in that with $K = K_0$, $\|F_k\|_{sp,22}$ [see also (2.15a)] is minimized over all stabilizing K for some $P \in R_p$ with $P_{22} = G_0$ and K_0 stabilizing P . Likewise, for the same P , $\|F_Q\|_{sp,22}$ [see also (2.15b)] is minimized over $Q \in RH^\infty$ with $Q = 0$. Necessary conditions on P

are that T defined in (2.16) satisfy

$$T(P) \in RH^\infty. \quad (2.23)$$

Sufficient conditions on P are that for arbitrary P_{12} , P_{21} satisfying $P_{12}M_0$, $\tilde{M}_0 P_{21} \in RH^\infty$, then $P_{11} = I - P_{12}M_0 \tilde{U}_0 P_{21}$. Moreover, this P [which satisfies (2.23)] achieves $e(z) = w(z)$.

Proof. See Appendix.

Remark. Notice that a P selection satisfying (2.23) may not lead to meaningful disturbance rejection indices. Take for example the case when $P_{12} = 0$ and $P_{11} = 0$. Here $e_k = 0$ irrespective of K_0 . It clearly makes sense to proceed with an index (2.22) that has meaning in an "engineering" sense.

Disturbance estimation

In order to minimize the effect of an unknown disturbance signal, it is reasonable to first estimate the disturbance, or at least its innovations, and then to feed back (somehow) such an estimate into the control signal. Of course, if a component or linear combination of the disturbance has no effect on the disturbance response, then there is no point in estimating this component.

Lemma 5. Consider the disturbance rejection controllers of Fig. 4 with (2.16), (2.23) holding. Sufficient conditions for w_k to be reconstructed from r_k by stable transfer functions are that

$$P_{21}^{-L} \text{ (left inverse) exists, } T_{21}^{-L} = P_{21}^{-L} \tilde{M}_0^{-1} \in RH^\infty. \quad (2.24)$$

Thus, ignoring initial conditions, under (2.24),

$$r(z) = T_{21}(z)w(z), \quad w(z) = T_{21}^{-L}(z)r(z). \quad (2.25)$$

For the case $P_{21} = \tilde{M}_0^{-1}$, then $e(z) = w(z) = r(z)$ (refer to Fig. 4).

Proof. See Appendix.

Remarks. The sufficient condition (2.32) is the simplest to work with. Others can be generated, for example, with $Yw(z) \in RH^\infty$ where $Y \in RH^\infty$ (2.32) can be replaced by

$$\begin{bmatrix} T_{21} \\ Y \end{bmatrix}^{-L} \in RH^\infty, \quad \hat{w}(z) = \begin{bmatrix} T_{21} \\ Y \end{bmatrix}^{-L} \begin{bmatrix} r(z) \\ 0 \end{bmatrix}.$$

This condition could be useful to apply when $w(z)$ is a deterministic signal, as a constant bias, or a finite sum of sinusoids of known period. Another sufficient condition is that for some W

such that $[T_{21} \ W] \in RH^+$ then

$$[T_{21} \ W]^{-L} \in RH^+, \quad \hat{v}(z) = [T_{21} \ W]^{-1} r(z)$$

where w_k is interpreted as being generated from a signal v_k driving a system with transfer function W . This condition could be useful when T_{21} is non-minimum phase and v_k is viewed as the innovations of w_k , and W is all pass.

Disturbance rejection controller for perturbed plant

Given that $K_0 \in R_p$ is an optimal disturbance rejection controller for $P \in R_p$ with $P_{22} = G_0$, then the disturbance response vector $e(z)$ is minimized in an L_2 sense over all $Q \in RH^+$ by the value $Q = 0$. Consider now the case of a plant $G \in R_p$, not the nominal one, but with the same disturbance rejection index as above. Let us consider an optimization over a restricted dimensional stabilizing proper Q_n , parametrized in terms of θ , a finite dimensional vector (matrix).

With P_{11} , P_{12} , P_{21} appropriately selected and $P_{22} = G$ (not G_0), consider the minimization task, in obvious notation

$$\min_{\text{stabilizing } \theta} \| [F_{Q_n}[P(G), K_0]]_s \|_2. \quad (2.26)$$

Of course such an optimization, even with F_Q affine in Q , is known to be difficult to carry out off-line with known G , and is impossible with G unknown. There may well be local minima which cause problems. We do not present any characterization of performance improvement for such an optimization since this appears too formidable a task. Suffice to say that the more restrictions placed on Q_n , the less is the potential for improvement, although it is known that the dimension on Q for optimum improvement is $\leq 2n - 1$ where n is the order of G_0 .

In the next section, we seek an on-line optimization procedure with this objective (3.26) in mind. Also, given that G is unknown, the selection of P_{11} , P_{12} , P_{21} as in examples (i) and (ii) would lead to unknown P_{11} , P_{12} , P_{21} and, therefore, a nonmeasurable disturbance response e_k . However, if the disturbance response, e_k is based on some measurable system signal such as the output weighted LQG scheme, we have a simple feasible formulation. Details are provided in the next section. On the other hand if the disturbance response is not based on a measurable system signal, P will have to be estimated on-line. We will only consider here the case when w , u are scalar to avoid parametrization uniqueness issues, complex formulations and the like which arise in dealing with the multivariable case. We define Q_n in terms of a

scalar Q under (2.24) as follows:

$$Q_n(z) = \frac{\beta_0 + \beta_1 z^{-1} + \dots + \beta_m z^{-m}}{1 + \alpha_1 z^{-1} + \dots + \alpha_n z^{-n}}$$

$$\theta' = [\alpha_1, \alpha_2, \dots, \alpha_n, \beta_0, \beta_1, \dots, \beta_m]. \quad (2.27)$$

3 ADAPTIVE- Q DISTURBANCE ESTIMATE FEEDBACK (DEF)

In this section, we introduce a direct adaptive DEF control scheme to perform on-line optimization of disturbance rejection indices. The adaptive loop feeds back disturbance estimates. It augments an existing controller interpreted as an optimal disturbance rejection controller. Attention is restricted to scalar variable nominal plants, although many of the concepts carry over in the multivariable case. The point in our derivations where we specifically require scalar plants is highlighted.

Consider some plant with Z -transform function $G \in R_p$ with a nominal representation $G_0 \in R_p$ and inputs u_k , outputs y_k as in (2.1) and (2.2). Consider also some stabilizing controller design $K_0 \in R_p$ for G_0 . Our objective is to enhance this controller by appropriate pre-processing of signals and introduction of an adaptive loop. We proceed by stages.

K_0 interpreted as a disturbance rejection controller

Perhaps the controller K_0 is known to be optimal for some known augmented plant P with $P_{22} = G_0$ in a disturbance rejection sense defined in Section 2. If P is unknown *a priori*, then we propose to select a P according to Lemma 3 such that (2.31) is satisfied. Then K_0 is optimal for this P , according to Lemma 3. Other ways of selecting P to satisfy different control objectives are presented later in the section and are investigated in greater detail in a companion paper (Moore and Tay, 1989a).

First stage of preprocessing

Let us augment the controller K_0 to yield $J_0 \in R_p$ as in (2.10) based on factorizations of G_0 , K_0 from (2.3) and (2.5). Two cases are of interest. The first is when we are free to construct the augmented controller J_0 . Such augmentations do not introduce additional dynamics if K_0 is a state estimate feedback design. Otherwise, the order of J_0 is at least that of the plant. This introduces an additional control variable s_k and measurement variable r_k as in Fig. 2a. Ignoring initial conditions, we have

$$\begin{bmatrix} u(z) \\ r(z) \end{bmatrix} = J_0 \begin{bmatrix} y(z) \\ s(z) \end{bmatrix}.$$

is scalar, (3.6) can be conveniently written as

$$\begin{aligned}\zeta_k &= \varphi'_k \theta + e_{k/\theta} \\ \varphi'_k &= [(e_{k-1} - \zeta_{k-1}) \cdots (e_{k-n} - \zeta_{k-n}) \\ &\quad - \xi_k - \xi_{k-1} \cdots - \xi_{k-m}].\end{aligned}\quad (3.7)$$

Proof. The proof of (3.5) and (3.6) is a trivial generalization of Lemma 3. The derivation of (3.7) is as follows. Let $\theta' = [\alpha_1 \cdots \alpha_n \beta_0 \cdots \beta_m]$. Noting that $s_{k-i/\theta} = s_{k-i}$, for $i \geq 1$, we have

$$\begin{aligned}s_{k/\theta} &= -\alpha_1 s_{k-1} - \cdots - \alpha_n s_{k-n} \\ &\quad + \beta_0 r_k + \cdots + \beta_m r_{k-m}.\end{aligned}\quad (3.8)$$

This gives

$$\begin{aligned}e_{k/\theta} &= \zeta_k + [-P_{12}M_0 s_{k-1}, \dots, -P_{12}M_0 s_{k-n}, \\ &\quad P_{12}M_0 \beta_0 r_k, \dots, P_{12}M_0 \beta_m r_{k-m}]\theta\end{aligned}$$

and noting that $P_{12}M_0 s_{k-i} = e_{k-i} - \zeta_{k-i}$, for $i \geq 1$, (3.7) follows.

Remark (1). That $e_{k/\theta}$, ζ_k are affine in θ is a by-product of the fact that F_Q is affine in Q . This property is crucial to conveniently applying any optimization of θ , off-line or on-line. The lemma forms the foundation of our adaptive algorithm which selects $\hat{\theta}_k$ to minimize an L_2 norm on $e_{k/\theta}$. (2) With Q_k stabilizing, then the effect of initial conditions decays asymptotically and it is reasonable to ignore the effect of initial conditions.

Third stage of preprocessing

It may be that there is *a priori* information about the frequency band of plant perturbations or controller non-robustness. It then makes sense to adapt Q only within such filter bands. This suggests that the residual r be preprocessed by a band-pass filter to yield filtered estimates r' to be used in place of r in the adaptive scheme. Further details on this are not explored in this paper.

Least squares θ selection

The off-line minimization task (2.26) (with $F_Q \in R_{sp}$) when translated into the time domain is the least squares minimization

$$\min_{\text{stabilizing } \theta} E[e'_k e_k]. \quad (3.9a)$$

For on-line least squares, the appropriate index is

$$\min_{\text{stabilizing } \theta} E[e'_{k/\theta} e_{k/\theta}]. \quad (3.9b)$$

This optimization leads to parameters $\hat{\theta}_k$ to be applied in the adaptive loop (3.1) and (3.2). Given this objective, the associated least squares

algorithm to use is as follows:

$$\begin{aligned}\hat{\theta}_k &= \hat{\theta}_{k-1} + \hat{P}_k \hat{\varphi}_k e_{k/\theta}, \quad e_{k/\theta} = \zeta_k - \hat{\varphi}'_k \hat{\theta}_{k-1} \\ \hat{P}_k &= \left(\sum_{i=1}^k \hat{\varphi}_i \hat{\varphi}_i' \right)^{-1} \\ &= \hat{P}_{k-1} - \hat{P}_{k-1} \hat{\varphi}_k [I + \hat{\varphi}'_k \hat{P}_{k-1} \hat{\varphi}_k]^{-1} \hat{\varphi}'_k \hat{P}_{k-1}, \\ &\quad \text{suitably initialized} \\ \hat{\varphi}'_k &= [(e_{k-1/\theta} - \zeta_{k-1}) \cdots (e_{k-n/\theta} - \zeta_{k-n}) \\ &\quad - \xi_k \cdots - \xi_{k-m}], \quad e_{k/\theta} = \zeta_k - \hat{\varphi}'_k \hat{\theta}_k\end{aligned}$$

or

$$\hat{\varphi}_k = \varphi_k \text{ as in (3.7)}. \quad (3.10)$$

Should $Q(\hat{\theta}_k) \notin RH^+$, then $\hat{\theta}_k$ may be projected into a stability domain in that $Q(\hat{\theta}_k) \in RH^+$ holds. Resetting $\hat{\theta}_k$ to zero is one method to achieve this. Other methods involve reduction of the step size appropriately. Such projection is not necessary when Q_k is a moving average and $\hat{\theta}_k$ is in a bounded domain. Another advantage of working with FIR Q_k is that fast least squares schemes with good numerical properties can be applied. Details are omitted here.

The adaptive DEF controller

The organization of the adaptive DEF controller is depicted in Fig. 6. Observe that once the prefiltering is in place, the adaptation of $\hat{\theta}_k$ is via a standard least squares scheme. In implementing the adaptive- Q scheme, Remark (4) following Theorem 1 should be noted.

Selection of P_{12} independent of G

From the algorithm of (3.10), and updating the adaptive loop (3.1), (3.2), it is observed that the formulation requires the implementation of P_{12} together with the measurability of e_k . It is usual to have e_k derived by appropriate filtering of u_k , y_k such as $e'_k = [u'_k y'_k]$. However selections such as those given in Example (1) or (2) in the known nominal plant case will lead to P_{12} involving the nominal plant G_0 . In the case, as here, when the plant is $G \neq G_0$, such selections of P_{12} will be in terms of the unknown G . This is not useful for our purposes. Here we exploit the fact that the pair P_{11} , P_{12} are not uniquely defined for a particular e_k expressed in terms of u_k and y_k . Taking an arbitrary full column rank P_{12} selection independent of G , although leading to complex formulations for P_{11} , appears reasonable. Of course, for different P_{12} there will be different regression vector filtering. These observations, we believe are crucial to the success of our approach.

We propose the following two selections of P_{12} .

In the framework of Section 2, with $e_k = H(z)w_k$, we have from (2.14),

$$\begin{aligned} P_{11}w_k &= e_k - P_{12}u_k \\ &= [H(z) - P_{12}K(Q_k)(I - GK(Q_k))^{-1}P_{21}]w_k. \end{aligned} \quad (3.11)$$

For the case $e'_k = [u'_k \ y'_k]$, then a reasonable choice is as follows:

$$\begin{aligned} P_{11} &= \begin{bmatrix} 0 \\ (I - GK(Q_k))^{-1} - K(Q_k)(I - GK(Q_k))^{-1}P_{21} \end{bmatrix}, \\ P_{12} &= \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \end{aligned} \quad (3.12)$$

Although P_{11} , implicitly in (3.13) or explicitly in (3.14), is non-linear and time-varying, it does not invalidate the derivation of Sections 2 and 3 since the derivations are carried through with $P_{11}w_k$ as a separate additional term. Note here that P_{11} is not implemented in the adaptive algorithm and, in general, knowledge of its form as given by (3.14) is not required. Another interesting selection of P_{12} is

$$P_{12} = M_0^{-1}. \quad (3.13)$$

Note that, for this selection of P_{12} , the implemented filters of Fig. 6a, b are simplified.

4. ANALYSIS

In this section, we study certain robustness results for the nominal controller with (non-adaptive) DEF controllers and stability/ convergence properties for the adaptive DEF controllers.

Robustness of DEF controllers

Theorem 1. Suppose K_0 is stabilizing for the plant G_0 with factorizations (2.3)–(2.6). Define corresponding factorizations for G_1 associated with the controller K_0 , which does not necessarily stabilize G_1 . Thus $G_1 = N_1M_1^{-1} = \tilde{M}_1^{-1}\tilde{N}_1$ where $N_1, M_1, \tilde{N}_1, \tilde{M}_1$ are not necessarily stable or coprime. Denote the class of all stabilizing controllers for G_0 characterized in terms of Q as in (2.8), (2.9) as $K(Q)$. Define $S = \tilde{M}_1N_0 - \tilde{N}_1M_0 = \tilde{M}_1(G_0 - G_1)M_0$, a frequency shaped version of $(G_0 - G_1)$. Then the necessary and sufficient conditions for $K(Q)$ to internally stabilize both G_0, G_1 with well-posed control loops are that a stable proper Q stabilizes S , or equivalently,

$$\begin{bmatrix} I & -Q \\ -S & I \end{bmatrix}^{-1} \text{ exists and belongs to } RH^\infty, \quad Q \in RH^\infty. \quad (4.1)$$

Moreover, a sufficient condition is

$$\|Q\| < \|S\|^{-1}, \quad Q \in RH^\infty. \quad (4.2)$$

Proof. Consider the stabilization of G_1 by $K(Q)$, and focus on the relevant version of (2.4) in that K is replaced by $K(Q)$ and G_0 by G_1 . Straightforward manipulations using definitions (2.5) and (2.8) and the Bezout identity (2.6) give (see also Tay *et al.*, 1989)

$$\begin{aligned} & \begin{bmatrix} I & -K(Q) \\ -G_1 & I \end{bmatrix}^{-1} \\ &= \left\{ \begin{bmatrix} \tilde{V}^{-1} & 0 \\ 0 & \tilde{M}_1^{-1} \end{bmatrix} \begin{bmatrix} \tilde{V} & -\tilde{U} \\ -\tilde{N}_1 & \tilde{M}_1 \end{bmatrix} \right\}^{-1} \text{ via (2.8)} \\ &= \left\{ \begin{bmatrix} I & -Q \\ -S & I \end{bmatrix} \begin{bmatrix} \tilde{V}_0 & -\tilde{U}_0 \\ -\tilde{N}_0 & \tilde{M}_0 \end{bmatrix} \right\}^{-1} \\ & \quad \times \begin{bmatrix} \tilde{V} & 0 \\ 0 & \tilde{M}_1 \end{bmatrix} \text{ via (2.6), (2.8)} \\ &= \begin{bmatrix} \tilde{V}_0 & -\tilde{U}_0 \\ -\tilde{N}_0 & \tilde{M}_0 \end{bmatrix}^{-1} \begin{bmatrix} I & -Q \\ -S & I \end{bmatrix}^{-1} \\ & \quad \times \left\{ \begin{bmatrix} I & -Q \\ -S & I \end{bmatrix} \begin{bmatrix} \tilde{V}_0 & 0 \\ 0 & \tilde{M}_0 \end{bmatrix} + \begin{bmatrix} Q\tilde{N}_0 & Q\tilde{M}_0 \\ S\tilde{V}_0 & S\tilde{U}_0 \end{bmatrix} \right\} \text{ via (2.8)} \\ &= \begin{bmatrix} I & -K_0 \\ -G_0 & I \end{bmatrix}^{-1} + \begin{bmatrix} M_0 & U_0 \\ N_0 & V_0 \end{bmatrix} \begin{bmatrix} I & -Q \\ -S & I \end{bmatrix}^{-1} \\ & \quad \times \begin{bmatrix} 0 & Q \\ S & 0 \end{bmatrix} \begin{bmatrix} \tilde{V}_0 & \tilde{U}_0 \\ \tilde{N}_0 & \tilde{M}_0 \end{bmatrix} \text{ via (2.5)} \\ &= \begin{bmatrix} I & -K_0 \\ -G_0 & I \end{bmatrix}^{-1} + \begin{bmatrix} M_0 & U_0 \\ N_0 & V_0 \end{bmatrix} \\ & \quad \times \left\{ \begin{bmatrix} I & -Q \\ -S & I \end{bmatrix}^{-1} - I \right\} \begin{bmatrix} \tilde{V}_0 & \tilde{U}_0 \\ \tilde{N}_0 & \tilde{M}_0 \end{bmatrix}. \end{aligned} \quad (4.3)$$

Sufficiency of the theorem is immediate. Equation (4.3) can be rewritten via (2.6) as

$$\begin{aligned} & \begin{bmatrix} I & -Q \\ -S & I \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{V}_0 & -\tilde{U}_0 \\ -\tilde{N}_0 & \tilde{M}_0 \end{bmatrix} \\ & \quad \times \left\{ \begin{bmatrix} I & -K(Q) \\ -G_1 & I \end{bmatrix}^{-1} - \begin{bmatrix} I & -K_0 \\ -G_0 & I \end{bmatrix}^{-1} \right\} \\ & \quad \times \begin{bmatrix} M_0 & -U_0 \\ -N_0 & V_0 \end{bmatrix} + I \end{aligned}$$

and necessity is established. Applying the small gain theorem (Desoer and Vidyasagar, 1975), $(I - SQ)^{-1} \in RH^\infty$ if $\|SQ\| < 1$, giving the sufficient condition (4.2) as claimed. $\Delta\Delta\Delta$

Remark 3. To complement the theorem, we note from Vidyasagar (1985) the parity interlacing requirement for simultaneous stabilization: Denote the non-minimum phase zeros, including those at infinity, of S in increasing order of magnitude as $\sigma_1, \sigma_2, \dots, \sigma_m$.

Then S is stabilizable by a stable compensator, Q , if the number of poles between any adjacent non-minimum phase zeros, (σ_i, σ_{i+1}) is even.

Remark 4. It turns out that if $K(Q)$ is implemented as a feedback loop consisting of J_0 and Q , then there is no additional requirement for stability. Details will be given in a subsequent paper. The message however is clear, namely that in implementing the adaptive Q schemes, $K(\hat{Q}_k)$ can be implemented either as one block, or as a J_0 block with an external feedback loop \hat{Q}_k .

Stability of the adaptive DEF controller

It is shown in Rohrs *et al.* (1985) that unmodelled dynamics or even small bounded disturbances can cause many adaptive control algorithms to go unstable. However, it is also shown in Chen and Guo (1988), Kresisselmeir and Anderson (1986) and Ioannou and Tsakalis (1986) that appropriate simple modifications or precautions are all that are needed to ensure robust stability. Such modifications are not discussed further in the present paper. Likewise a complete stability analysis is beyond the scope of this paper.

We stress that the proposed adaptive scheme is not presented as a globally stable scheme in the sense of the relatively simple arrangement as presented in Chen and Guo (1988), Kresisselmeir and Anderson (1986) and Ioannou and Tsakalis (1986). At most, local stability is claimed as for the schemes therein, when some of their assumptions are relaxed. Rather, the scheme proposed is presented as one rational approach for adaptive augmentation of existing robust fixed controllers aimed to enhance the performance of such controllers. The limited objective is to improve performance for plants in the neighborhood of the nominal plant used in the robust fixed controller design. It must be admitted that, although the prime engineering objective is to enhance performance at the margins of acceptability of the robust design; ironically, this is the case least subject to analysis, being usually outside a region for which even local stability results can be guaranteed.

Convergence results for adaptive DEF controller

The adaptive DEF controller is constructed with the objective of evanescent for the case when the plant is the nominal one and the controller has optimal performance and enhancing performance otherwise. Does it achieve these objectives? We present convergence results based on the assumption that for small perturbations of the nominal model, the

closed-loop adaptive controller with $\|Q\|$ suitably constrained is stabilizing. This assumption enables us to use the ODE approach in our analysis.

Nominal plant case

Theorem 2 Part (i). Consider the adaptive DEF controller of Sections 2 and 3 applied to a nominal plant G_0 . If $\hat{\theta}_k$ converges to some value θ^* with $Q_\theta^* \in RH^+$, then the DEF controller $K(\theta^*)$ is stabilizing for G_0 and optimum in the minimum variance sense of (3.9), or, equivalently, the disturbance rejection sense of (2.22b). Moreover, when $K_0 = K(\theta = 0)$ is a unique optimum disturbance rejection controller, then if $\hat{\theta}_k$ converges, it converges to $\theta^* = 0$.

Part (ii). With plant disturbance such that ξ_k is persistently exciting for each fixed Q , in that $E[\hat{q}_k \hat{q}_k^T] > 0$, then $\hat{\theta}_k$ of the least squares algorithm converges almost surely and asymptotic optimality is achieved as in Part (i).

Proof Part (i). The first result is immediate since the DEF controller is stabilizing for all fixed θ with $Q_\theta \in RH^+$. Now since $\hat{\theta}_k$ is a least squares estimate, it minimizes the index $k^{-1} \sum_1^k \|\xi_k - \theta' q_k\|^2$ for all k . With $\hat{\theta}_k$ converging to θ^* , and closed loop stability guaranteed since $Q_\theta^* \in RH^+$, then as $k \rightarrow \infty$, this least squares index approaches the index $E \|\xi_k - \theta'^* q_k\|^2 = E[\hat{e}_k^T \hat{e}_k]$. (This is a well known asymptotic ergodicity result.) Consequently, θ^* optimizes the index (3.6). Moreover, when K_0 is a unique optimum controller, there is a contradiction unless $\theta^* = 0$, since $K(\theta)$ is uniquely parametrized in terms of θ .

Proof Part (ii). First observe that for "frozen" θ , $\hat{q}_k = \partial \hat{e}_k / \partial \theta$, due to the fact that e_k is affine in Q . Consequently, the least squares algorithm is a standard recursive prediction error (RPE) algorithm, with associated ordinary differential equation (ODE)

$$\begin{aligned} \dot{\theta}_r &= R_r^{-1} f(\theta_r), \quad \dot{R}_r = G(\theta_r) - R_r, \\ f(\theta_r) &= -E[\hat{q}_k \hat{e}_k], \quad G(\theta_r) = E[\hat{q}_k \hat{q}_k^T]. \end{aligned} \quad (4.4)$$

Its Lyapunov function is $V(\theta) = 1/2E[\hat{e}_k^T \hat{e}_k] \geq 0$ with $\dot{V}(\theta) = \partial V / \partial \theta$ and $\dot{\theta} = -f'(\theta_r) R_r^{-1} f(\theta_r) \leq 0$. Under excitation of \hat{q}_k so that $R > 0$, then the convergence of V ensures that θ_r of the ODE globally converges to the set $\{\theta | f(\theta) = 0\}$. Since the theory for the class of all stabilizing controllers reviewed in Section 2, tells us that the closed loop system for all $\hat{\theta}_k$ is projected into a stability domain, the ODE theory of Ljung (1977) now tells us that $\hat{\theta}_k$ itself

converges. The results of Part (i) apply to ensure asymptotic optimality.

Remark. The key convergence condition above is that $\hat{\varphi}_k$ for each frozen θ be persistently exciting. Such can be achieved when w_k is sufficiently rich as when its innovations are white with a variance bounded below and $\hat{\varphi}_k$ is reachable from w_k ; see Moore (1987) and Green and Moore (1987) for details. Sufficient conditions are that $r_k = \tilde{M}_0 y_k - \tilde{N}_0 u_k$ be full rank (this is also necessary), and that $Q(z)$ is minimal in that there are no pole/zero cancellations. This latter condition on Q can be by-passed by a mild modification to the least squares algorithm to ensure that $P_k^{-1} > k \delta I$ for some "small" $\delta > 0$. Then in (4.3), $\hat{R} = G - R + \delta I$ and $R > 0$, irrespective of excitation of $\hat{\varphi}_k$.

Non-nominal plant

In the previous section, Lemma 6 was used to generate the adaptive algorithm. The algorithm justified as a least squares scheme for the nominal case is in fact a recursive prediction error (RPE) scheme for plants ($P_{22} \neq G_0$) other than the nominal one. This is summarized as follows.

Lemma 7. The adaptive DEF controller of Sections 2 and 3 applied to any plant G is a recursive prediction error algorithm (RPE) in that

$$\hat{\varphi}_k = -\partial e_k(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{k-2}, \theta) / \partial \theta|_{\theta = \hat{\theta}_{k-1}} \quad (4.5)$$

Consider that the controller $K(\theta)$ with $Q = 0$ is an optimal stabilizing disturbance rejection controller for a nominal plant G_0 , and is also stabilizing for the actual plant G . Consider that the parameter update equations (3.10) initialized by $\hat{\theta}_0 = 0$ are modified by stepsize reductions to ensure that $\|\hat{\theta}_k\| < \delta$ for δ suitably small. Then the DEF/RPE scheme is globally convergent to the optimal (off-line) prediction error controller set $\{\theta^0 \mid \|\theta^0\| < \delta\}$.

Proof. That $\hat{\varphi}_k$ is dependent on $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{k-1}$ can be observed from (3.4) since u_k is generated from $\hat{\theta}$ up to $\hat{\theta}_{k-1}$. From (3.7), (4.5) which is the defining property of an RPE scheme follows. The rest of the Lemma follows from standard properties of an RPE scheme with projection into a stability domain.

Remark 5. The selection of a value of δ can be guided by Theorem 1. Should $\hat{\theta}_k$ be such that, for each k , the closed loop system is

asymptotically stable, then the projection into $\|\hat{\theta}_k\| < \delta$ is not required. This projection is after all a projection into a stability domain. Of course for some arbitrary plant G , there is no way of checking on line whether $\hat{\theta}_k$ lies in a stability domain, and thus the theorem is useful only when G is in the neighborhood of G_0 . Of course, as subsequent simulations show, the DEF/RPE scheme may converge even when the conservative conditions of the above theorem are violated, but precise convergence results for these cases are not yet available.

Observe that there is no positive real condition on the plant noise as in extended least squares algorithms. Here, the RPE algorithm "substitutes" a projection into a stability domain condition for θ , which is automatic when Q is forced to be stable as when Q is a moving average operator.

5. SIMULATION RESULTS

Here simulations of the adaptive- Q controller are presented to illustrate their performance enhancement capabilities.

Example (3). Enhancement of an LQG Controller.

Consider a nominal plant (2.11) parametrized as

$$A_0 = \begin{bmatrix} 3.0 & 1 \\ -3.25 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 1 \\ -1.1 \end{bmatrix}, \quad C_0 = [1 \quad 0], \quad (5.1)$$

$$D_0 = [0], \quad \Gamma_0 = \begin{bmatrix} 3.2 \\ -2.75 \end{bmatrix}$$

with unstable poles at $z = 1.5 \pm j1.0$ and a non-minimum phase zero at $z = 1.1$. Consider the perturbed plant

$$A_1 = \begin{bmatrix} 2.8 & 1 \\ -2.96 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 1 \\ -1.3 \end{bmatrix}, \quad C_1 = [1 \quad 0], \quad (5.2)$$

$$D_1 = [0], \quad \Gamma_1 = \begin{bmatrix} 2.81 \\ -2.83 \end{bmatrix}$$

which has unstable poles at $z = 1.6 \pm j1.0$ and a non-minimum phase zero at $z = -1.3$. Consider also a second perturbed plant

$$A_2 = \begin{bmatrix} 3.6 & 1 \\ -5.49 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 \\ -1.5 \end{bmatrix}, \quad C_2 = [1 \quad 0], \quad (5.3)$$

$$D_2 = [0], \quad \Gamma_2 = \begin{bmatrix} 3.61 \\ -5.36 \end{bmatrix}$$

which has unstable poles at $z = 1.8 \pm j1.5$ and a non-minimum phase zero at $z = -1.5$. An LQG controller is designed based on the index (2.22) with $Q_c = C_0' C_0$ and $R_c = 0.2704$. The factorization of (2.7b) is used to construct J_0 via (2.13).

TABLE 1

	Nominal plant	Plant	Plant (ii)
J_{LOG} with fixed nominal controller	3.99	5.449	unstable
J_{LOG} with nominal controller and an adapted 3-term Q	4.00	4.005	9.649
J_{LOG} with ideal fixed controller	3.99	3.671	7.809

The nominal controller stabilizes the first perturbed plant model with closed loop eigenvalues of $0.513 \pm j0.3672$ and $-0.5318 \pm j0.0478$. It however fails to stabilize the second plant model. The closed loop eigenvalues are $-0.6955 \pm j0.78$, 0.5048 and 0.0350 .

Applying the DEF adaptive controller of Sections 2 and 3 discussed in Example (i) with a three-term scalar moving average (MA) Q_k adapted on-line, Table 1 summarizes the results for simulation of 5000 sample points when w_k is a white noise, zero mean sequence of variance one.

In both simulations, the performance is shown to have improved with an adapted Q_k . For plant (ii), besides an improvement in performance, the scheme enhanced the overall robustness of the controller. Figure 7 shows plots of the plant output, input and estimates of θ . The estimates converge after about 40 samples and K_n together with the augmentation Q_k , stabilizes the plant.

Example 4. Enhancement of a one-horizon LQ controller.

The one-horizon LQ controller is a suboptimal scheme in relation to an infinite LQ controller.

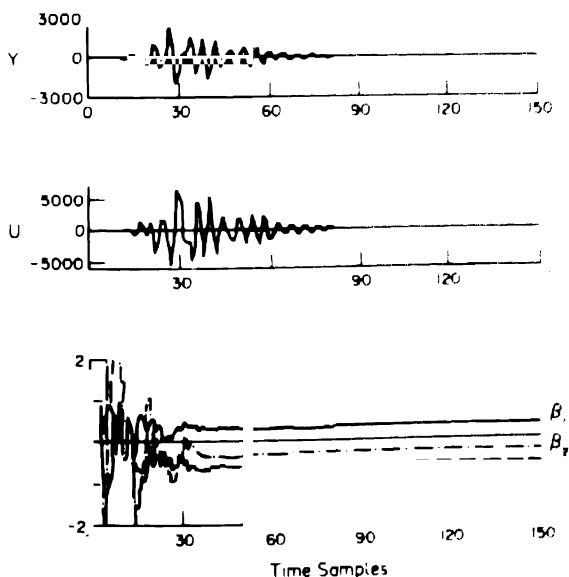


FIG. 7. Simulation example (i), plant (ii). Top: Plant output, centre: Plant input, bottom: Parameter estimates of Q .

Its performance can be enhanced by an on-line adaptive- Q algorithm developed in relation to Example (2) of earlier sections. Simulations on the nominal plant and index of Example (3) show that the infinite horizon index improves from 5.18 for the optimal one-horizon LQ controller to 4.48 when augmented with the adaptive- Q controller.

6. CONCLUSIONS

A disturbance estimate feedback (DEF) adaptive- Q control scheme using least squares parameter updates is described. The adaptations constitute a recursive prediction error (RPE) scheme and inherit the stability properties of such. The scheme can, if required, limit its search to the space of all stabilizing linear proper controllers for a nominal linear plant model of the actual plant. Such schemes are conservative and can be viewed as seeking on-line a robust controller for the nominal plant. The scheme minimizes the effect of disturbances based on the same performance criterion as that of the design for the nominal fixed controller, thereby improving the performance over that of the fixed controller and yet preserving the original objective of the fixed controller design. The adaptive loop evanesces when the plant is the nominal one and the fixed controller optimal, but otherwise serves to improve the fixed controller. Simulation results verify the effectiveness of the adaptive scheme. For the more radical versions of the DEF scheme, the adaptations need not limit themselves to being stabilizing for the nominal plant while adapting. Then the potential for stabilizing a wider class of plants exists, but such a potential would need to be carefully explored before application.

Related research is to apply the techniques of this paper to adaptive two degree-of-freedom tracking schemes as in Tay and Moore (1990), and also to the situation where the robust fixed controller of this paper is also adaptive, leading to an adaptive scheme which combines the advantages of indirect and direct adaptive control [see also an adaptive frequency shaped Kalman Filter for improved performance in unknown colored noise environments (Moore and Tay, 1989b)]. Simulation studies support the power of these approaches.

Acknowledgement—This work was partially supported by DSTO Australia and Boeing.

REFERENCES

- Anderson, B. D. O., J. B. Moore and R. M. Hawkes (1978). Model approximations via prediction error identification. *Automatica*, **14**, 615–622.
- Chakravarty, A. and J. B. Moore (1986). Aircraft flutter

- suppression via adaptive LQG control. *Proc Amer Control Conf*, pp. 488–493.
- Chen, H. F. and L. Guo (1988). A robust stochastic adaptive controller. *IEEE Trans. Aut. Control*, **AC-33**, 1035–1043.
- Desoer, C. A. and M. Vidyasagar (1975). *Feedback Systems: Input-Output Properties*. Academic Press, New York.
- Goodwin, G. C., P. J. Ramadge and P. E. Caines (1980). Discrete time multivariable adaptive control. *IEEE Trans. Aut. Control*, **AC-25**, 449–456.
- Green, M. and J. B. Moore (1987). Persistency of excitation in linear systems. *Syst. Control Lett.* See also an extended version in *Proc. Amer. Control Conf.*, Boston, pp. 412–417, June 1985.
- Ionnou, P. and K. S. Tsakalis (1986). A robust direct adaptive controller. *IEEE Trans. Aut. Control*, **AC-31**.
- Kosut, R. L. and B. D. O. Anderson (1984). Robust adaptive control: Condition for local stability. *Proc. 23rd IEEE Conf. on Decision and Control*, Las Vegas, NV.
- Kreisselmeier, G. and B. D. O. Anderson (1986). Robust model reference adaptive control. *IEEE Trans. Aut. Control*, **AC-31**, 127–133.
- Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Aut. Control*, **AC-22**, 551–575.
- Moore, J. B. (1985). Stochastic adaptive control via consistent parameter estimation. *Proc. IFAC Identification Conference*, York, U.K. pp. 611–616.
- Moore, J. B. (1987). A universality advantage of stochastic excitation signals for adaptive control. *Syst. Control Lett.*, **9**, 55–58.
- Moore, J. B., K. Glover and A. Telford (1990). All stabilizing controllers as frequency shaped state estimate feedback. *IEEE Trans. Aut. Control*, **AC-35**, 203–208.
- Moore, J. B., A. T. Holtz and D. Gangsaas (1982). Adaptive flutter suppression as a complement to LQG based aircraft control. *Proc. 6th Identification Conference*, Washington, DC.
- Moore, J. B. and T. T. Tay (1989). Adaptive control within the class of stabilizing controllers for a time-varying nominal plant. *Int. J. Control*, **50**, 33–53.
- Moore, J. B. and T. T. Tay (1989b). Adaptive frequency shaped kalman filter. *IEEE Trans. Aut. Control*, **AC-34**, 231–236. See also *Proc. 1st LASED, Int. Symp. on Signal Processing and its Applications*, Brisbane, August 1987.
- Narendra, K. S., Y. H. Lin and L. S. Valavani (1980). Stable adaptive controller design. Part II: Proof of stability. *IEEE Trans. Aut. Control*, **AC-25**, p. 440.
- Rohrs, C. E., L. Valavani, M. Athans and G. Stein (1985). Robustness of continuous time adaptive control algorithms in the presence of unmodeled dynamics. *IEEE Trans. Aut. Control*, **AC-30**, 881–889.
- Tay, T. T. and J. B. Moore (1990). Performance enhancement of two degree-of-freedom controllers via adaptive techniques. *J. Adaptive Control Signal Process.*, **4**, 69–84.
- Tay, T. T., J. B. Moore and R. Horowitz (1989). Indirect adaptive techniques for fixed controller performance enhancement. *Int. J. Control*, **50**, 1941–1960.
- Vidyasagar, M. (1985). *Control System Synthesis: A Factorization Approach*. MIT Press, Cambridge, MA, U.S.A.

APPENDIX

Verification of (2.6) under (2.7)

Recalling that

$$\begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} = \begin{bmatrix} A_1 & B_1 C_2 & B_1 D_2 \\ 0 & A_2 & B_2 \\ C_1 & D_1 C_2 & D_1 D_2 \end{bmatrix} \quad (A.1)$$

Then for the second set of factorizations

$$\begin{bmatrix} A + ALC & B - AL \\ F + FLC & I - FL \\ C & 0 & I \end{bmatrix} \begin{bmatrix} A + BF & B & -(A + BF)L \\ F & I & -FL \\ C & 0 & I \end{bmatrix} =$$

$$= \begin{bmatrix} A + ALC & ALC - BF & -B & (A + BF)L \\ 0 & A + BF & B & -(A + BF)L \\ F + FLC & F + FLC & I & 0 \\ C & C & 0 & I \end{bmatrix}_1$$

$$= \begin{bmatrix} A + BF & ALC - BF & -B & (A + BF)L \\ 0 & A + ALC & 0 & 0 \\ 0 & F + FLC & I & 0 \\ 0 & C & 0 & I \end{bmatrix}_1$$

$$= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \quad (A.2)$$

The second equality follows from a co-ordinate basis change, and the third by removal of uncontrollable and unobservable modes. The verification of (2.6a) follows likewise.

Proof of Lemma 2.1 From Fig. 2b, the relationship between e and u is given by

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = \begin{bmatrix} I - S_{12} & 0 \\ 0 & I - Q \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} S_{11} & 0 & 0 \\ 0 & I & 0 \\ S_{21} & 0 & I \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \quad (A.3)$$

$$e = \begin{bmatrix} S_{11} + S_{12}QS_{21} & S_{12} & S_{12}Q \\ QS_{21} & I & Q \\ S_{21} & 0 & I \end{bmatrix}$$

and (2.12) follows.

Since W is affine in Q with the "coefficients" involving S_{ij} belonging to RH^+ via (2.10b), then $Q \in RH^+$ implies $W \in RH^+$. Notice also that $Q \notin RH^+$ implies that $W_{12} \notin RH^+$ and thus $W \notin RH^+$. Thus $W \in RH^+$ implies $Q \in RH^+$. Now, it is known that internal (asymptotic) stability of the system of Fig. 2 under detectability and stabilizability of its blocks (G_0, J, Q) is equivalent to all the transfer function between the u_i and e_i being (asymptotically) stable. Thus internal stability (under the lemma condition) is equivalent to W being (asymptotically) stable. As in the earlier reasoning, internal stability holds if and only if Q is (asymptotically) stable. The part (i) results of the lemma are now established. Likewise, since S_{ij} are bounded-input, bounded-output operators, under part (ii) assumptions, the part (ii) results follow.

Proof of Lemma 2. For the schemes of Fig. 4 with $K = K_0$, $Q = 0$, internal stability is equivalent to the following condition on input/output transfer functions.

Since $V_0, U_0 \in RH^+$ are coprime, and $V_0, U_0 \in RH^+$, then

$$P_{12}M_0[V_0 \ U_0] \in RH^+ \Leftrightarrow P_{12}M_0 \in RH^+ \quad (A.4)$$

$$\begin{bmatrix} U_0 \\ V_0 \end{bmatrix} M_0 P_{12} \in RH^+ \Leftrightarrow M_0 P_{12} \in RH^+$$

and (A.5) is equivalent to (2.31). The lemma results are now immediate.

Proof of Lemma 3. Writing

$$\begin{bmatrix} e(z) \\ v(z) \end{bmatrix} = P(G) \begin{bmatrix} w(z) \\ u(z) \end{bmatrix}, \quad \begin{bmatrix} u(z) \\ v(z) \end{bmatrix} = J_0 \begin{bmatrix} w(z) \\ v(z) \end{bmatrix}$$

and noting

$$\begin{bmatrix} e(z) \\ v(z) \end{bmatrix} = T(G) \begin{bmatrix} w(z) \\ v(z) \end{bmatrix} \quad (A.5)$$

straightforward algebraic manipulations generalize T of (2.16) to $T(G)$ of (2.20) when P is generalized to $P(G)$.

Define coprime factorizations for G as $G = NM^{-1} = \tilde{M}^{-1}\tilde{N}$ such that with the factorizations of K_0 as in (2.5), the double Bezout identity of (2.6) is satisfied. Here $M, N, \tilde{M}, \tilde{N} \in RH^+$ only if K_0 stabilizes G . Now

$$e(z) = T_{11}(G)w(z) + T_{12}(G)Q(I - T_{22}(G)Q)^{-1}T_{21}(G)w(z). \quad (A.6)$$

Substituting for $T_{22}(G)$ from (2.20) and applying the Bezout identity,

$$\begin{aligned} (I - T_{22}(G)Q)^{-1} &= [I - V_0^{-1}(I - GK_0)^{-1}GV_0^{-1}Q \\ &\quad + V_0^{-1}N_0Q] \\ &= [V_0 - GV_0 - GV_0^{-1}Q + (I - GK_0)N_0Q]^{-1} \\ &\quad + (I - GK_0)V_0 \\ &= [V_0 + N_0Q - GV_0 - GV_0^{-1}(I + U_0N_0)Q]^{-1} \\ &\quad \times (I - GK_0)V_0 \\ &= (V - GU)^{-1}(I - GK_0)V_0 \\ &= (\tilde{M}V - \tilde{N}U)^{-1} \end{aligned} \quad (A.7)$$

Thus (A.6) can be re-organized, via (2.20) and (2.8) as

$$\begin{aligned} e(z) &= [P_{11} + P_{12}U_0\tilde{M}]w(z) + P_{12}MQ \\ &\quad \times (\tilde{M}V - \tilde{N}U)^{-1}MP_{21}w(z) \\ &= P_{11}w(z) + P_{12}[U_0(\tilde{M}V - \tilde{N}U) + MQ] \\ &\quad \times (\tilde{M}V - \tilde{N}U)^{-1}MP_{21}w(z) \end{aligned} \quad (A.8)$$

Now the square bracket term of (A.8) can be re-organized as

$$\begin{aligned} [U_0\tilde{M}(V_0 + N_0Q) + U_0N(U_0 + M_0Q) + MQ] \\ &= [U_0 + (U_0\tilde{M}N_0 - U_0NM_0)Q + MQ] \\ &= [U_0 + (MU_0N_0 - U_0NM_0)Q + MQ] \\ &= [U_0 + (MV_0 - U_0N)M_0Q] = [U_0 + M_0Q] \end{aligned}$$

giving a re-organization of (A.8) as, writing $V = V(Q)$, $U = U(Q)$,

$$\begin{aligned} e(z) &= P_{11}w(z) + P_{12}[U_0 + M_0Q] \\ &\quad \times [\tilde{M}V(Q) - \tilde{N}U(Q)]^{-1}MP_{21}w(z) \end{aligned} \quad (A.9)$$

From (2.10), we have $v(z) = V_0^{-1}v(z) + V_0^{-1}N_0v(z)$ and $w(z) = K_0v(z) + V_0^{-1}v(z)$. Applying the Bezout identity (2.6), gives $v(z) = M_0v(z) + N_0w(z)$. Now from

$$v(z) = Gu(z) + P_2w(z) = u(z) + K(Q)w(z),$$

so that

$$\begin{aligned} v(z) &= U_1(Q)[\tilde{M}V(Q) - \tilde{N}U(Q)]^{-1}MP_{21}w(z), \\ w(z) &= U_1(Q)[\tilde{M}V(Q) - \tilde{N}U(Q)]^{-1}MP_{21}w(z) \end{aligned}$$

giving

$$v(z) = M_0v(z) + N_0w(z) = [\tilde{M}V(Q) - \tilde{N}U(Q)]^{-1}MP_{21}w(z) \quad (A.10)$$

Substituting into (A.9) gives the lemma result (2.21).

Proof of Lemma 4. For arbitrary stabilizing $K_0 \in R_1$ for $G_0 \in R_{n_0}$ it is always possible to choose P of the form

$$P = \begin{bmatrix} I - P_{12}K_0(I - G_0K_0)^{-1}P_{11} & P_{11} \\ P_{12} & G_0 \end{bmatrix} \quad (A.11)$$

This gives $I_{K_0} = I_{G_0} = I$ so that $[I_{K_0}]_0 = 0$ and $e_k = w_k$. Therefore the index of (2.22) is minimized. Necessary conditions for (2.23) follow directly from Lemma 2.

Proof of Lemma 5. The assumption that γ_1 has full column rank guarantees the existence of P_1^{-1} and thus $I_{\gamma_1}^{-1}$. Moreover, (2.25) follows from the condition $I_{\gamma_1} = 0$ and a stability on $I_{\gamma_1}^{-1}$.

Continuous-time Generalized Predictive Control (CGPC)*

H. DEMIRCIOĞLU† and P. J. GAWTHROP‡

The well known and successful discrete-time Generalized Predictive Controller (GPC) is rederived in a continuous-time setting. It is shown that the resulting algorithm has all the power and flexibility of the GPC, without having the drawbacks of the discrete-time formulation.

Key Words—Adaptive control, control system design, optimal control, predictive control, self-tuning regulators

Abstract—A continuous-time version of the discrete-time Generalized Predictive Controller is presented. The continuous-time formulation arises from a mixture of two kinds of analogy between continuous and discrete-time systems: a *physical* analogy and an *algebraic* analogy. Emphasis is placed on the differences arising from a continuous-time formulation, and the relative merits of a continuous and a discrete-time approach are given. Although mainly concerned with the design algorithm itself, the paper also indicates how a self-tuning version can be implemented. Illustrative simulations are given.

1. INTRODUCTION

IN RECENT YEARS, long-range predictive control (LRPC) has attracted much attention as an underlying design method for self-tuning controllers (Richalet *et al.*, 1978; Cutler and Ramaker, 1980; de Keyser and van Cauwenberghe, 1981, 1985; Ydstie, 1984; Peterka, 1984; Mosca *et al.*, 1984; de Keyser *et al.*, 1988; Clarke and Zhang, 1987; Clarke *et al.*, 1987; Lelic and Zarrop, 1987). This is due to its superior robustness compared to some other self-tuning control methods such as GMV and Pole-placement; it owes this good robustness property to the minimization of a multi-step cost function.

The method can be summarized as follows:

1. Predict the system output over a range of future times.
2. Assuming that the future setpoint is known, choose a set of future controls which minimize the future errors between the predicted future output and the future setpoint.
3. Use the first element $u(t)$ as a current input and repeat the whole procedure at the next time instant, that is, use a receding horizon strategy.

All of the above algorithms were developed in a discrete-time context. In contrast, we propose a method developed in a continuous-time context. In particular, a *continuous-time* analogue of the *discrete-time* GPC (Generalized Predictive Control) proposed by Clarke *et al.* (1987) is developed; hence the name: Continuous-time Generalized Predictive Control (CGPC).

As will be shown, the CGPC has properties similar to those of the discrete-time GPC, albeit with a continuous-time interpretation. It has similar design parameters which have similar effects. The control weighting is not essential for the control of non-minimum phase systems. It can easily handle complex systems, such as at the same time open-loop unstable, non-minimum phase and higher order. It can control time delay systems without any problem.

The paper is organized as follows. The CGPC algorithm is introduced in Section 2. Section 3 examines the resulting closed-loop system in detail. In Section 4, the effects of CGPC parameters on the closed-loop system response are discussed. An illustrative simulation study is given in Section 5 and Section 6 concludes the paper.

* Received 6 October 1988; revised 5 July 1989; received in final form 15 February 1990. The original version of this paper was presented at the IFAC Symposium on Adaptive Systems in Control and Signal Processing which was held in Glasgow, Scotland, U.K. during April 1989. The published Proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor D. W. Clarke under the direction of Editor P. C. Parks.

† Department of Mechanical Engineering, The University, Glasgow G12 8QQ, U.K.

‡ Author to whom all correspondence should be addressed.

2 DEVELOPMENT OF THE CGPC ALGORITHM

2.1. System description

The following strictly proper single-input single-output system is used in the development of the algorithm.

$$Y(s) = \frac{B(s)}{A(s)} U(s) + \frac{C(s)}{A(s)} V(s) \quad (2.1.1)$$

where $A(s)$, $B(s)$ and $C(s)$ are polynomials in the Laplace operator s . $Y(s)$, $U(s)$ and $V(s)$ are the system output, control input and disturbance input respectively.

No special assumptions are placed on the disturbance $V(s)$ and thus the polynomial $C(s)$ will be considered as a *design* polynomial having all its roots in the left-half s plane. The degree of $C(s)$ depends on the characteristics of the disturbances. In many cases, the disturbance component of the system is such that we would not wish to differentiate it; this can be modelled by choosing $\deg(C(s)) = \deg(A(s)) - 1$. An even worse case occurs when we would not wish even to use the system output directly; this can be modelled by choosing $\deg(C(s)) = \deg(A(s))$. In the sequel it will be assumed that $\deg(C(s)) = \deg(A(s)) - 1$.

Note that equation 2.1.1 does not include any time delay explicitly. However, the time delay can be approximately incorporated into the A and B polynomials by using a suitable approximation such as the Padé approximation (Marshall, 1979; Gawthrop, 1987; de Souza *et al.*, 1988). Therefore, systems with time delay can also be modelled (approximately) by the same equation. In other words, instead of a system with time delay we can consider a higher-order system without time delay; this is the approach used here.

2.2. Derivative emulation

The notion of emulator was first introduced in Gawthrop (1986) to describe the dynamic systems which emulate unrealizable operations. Examples of such unrealizable operations in control systems design include: taking derivatives of the output, cancelling non-minimum phase zeros and removing time delay by an inverse delay (prediction). Control systems which have an emulator in the feedback loop will be called "*Emulator Based Control*" (EBC). It can be shown (Gawthrop, 1986, 1987) that the EBC corresponds to the discrete time GMV (Generalized Minimum Variance) method (Clarke and Gawthrop, 1975, 1979). More details of emulators and EBC can be found elsewhere (Gawthrop, 1987); here, we will only

consider the emulation of output derivatives relevant to the paper.

Taking the derivative (with respect to time) of a signal in time domain corresponds to multiplication by s in the Laplace domain (assuming zero initial conditions). Thus the k th derivative of the system output can be written in the Laplace domain as

$$Y_k(s) = s^k Y(s) = \frac{s^k B(s)}{A(s)} U(s) + \frac{s^k C(s)}{A(s)} V(s) \quad (2.2.1)$$

and the s^k -multiplied disturbance transfer function decomposed into two parts as

$$\frac{s^k C(s)}{A(s)} = E_k(s) + \frac{F_k(s)}{A(s)} \quad (2.2.2)$$

where†

$$\deg(F_k) = \deg(A) - 1$$

$$\deg(E_k) = k - 1.$$

The transfer function F_k/A represents the strictly proper part of $s^k C/A$ and E_k the improper remainder.

Using identity (2.2.2) $Y_k(s)$ may be written as the sum of an emulated value $Y_k^*(s)$ and the corresponding error $E_k^*(s)$

$$Y_k(s) = Y_k^*(s) + E_k^*(s) \quad (2.2.3)$$

where

$$Y_k^*(s) = \frac{s^k B}{A} U(s) + \frac{F_k}{A} V(s) \quad (2.2.4)$$

and

$$E_k^*(s) = E_k V(s). \quad (2.2.5)$$

Equation (2.2.4) cannot be implemented as $V(s)$ is unknown. But from the system equation (2.1.1)

$$V(s) = \frac{A}{C} Y(s) - \frac{B}{C} U(s). \quad (2.2.6)$$

Substituting equation (2.2.6) into equation (2.2.4) and using identity (2.2.2) the following expression for the emulated value of the k th derivative of the output arises.

$$Y_k^*(s) = \frac{E_k B}{C} U(s) + \frac{F_k}{C} Y(s). \quad (2.2.7)$$

Notice that the relative order of $E_k B/C$ is $\rho - k$, where ρ is the relative order of the system. For this term to be realisable, $k \leq \rho$. The transfer function F_k/C is proper.

The equations leading to equation (2.2.7) are *algebraically* equivalent to those leading to a discrete-time k -step ahead predictor, but the interpretation is different.

† In the rest of the paper, the argument of the polynomials will be dropped where there is no ambiguity.

2.3. Output prediction

As its name implies, predictive control is based on the prediction of the future output. The GPC is based on a *range* of future output predictions. This is in contrast to the other controllers based on output prediction, such as MV and GMV, in which a prediction for only *one* value of prediction time is needed. Thus the predictor within the CGPC is designed as a function of a variable time T into the future so that by varying T a range of output predictions can be obtained.

As discussed by Åström (1970) the derivation of a discrete-time predictor is straightforward due to the direct correspondence between the series expansion of a discrete-time transfer function and the associated impulse response. In other words, the discrete-time transform variable z corresponds to a forward time shift. In continuous-time, the task of designing a predictor is less obvious. One possible approach is to make use of the fact that the current derivatives of a continuous-time signal imply the future development of that signal. In other words, if the output derivatives of a continuous function at time t are known, it is possible to predict the future values. This T -ahead predictor is approximated by a truncated Maclaurin series as follows:[†]

$$\hat{y}(t+T) = y(t) + \sum_{k=1}^{N_p} y_k(t) \frac{T^k}{k!} \quad (2.3.1)$$

where

$$y_k(t) = \frac{d^k \hat{y}(t+T)}{dT^k} \text{ (at } T=0) = \frac{d^k y(t)}{dt^k} \quad (2.3.2)$$

N_p = predictor order.

An appropriate value for N_p is discussed in Section 4.3; but, roughly speaking, the larger T the larger N_p should be for a good prediction.

As stated above, the derivatives of the *current* system output (at time t) are needed to predict the *future* output. However, taking derivatives of the system output is not feasible because of noise amplification. The solution to this problem is to replace the output derivatives in equation (2.3.1) by their emulated values. As discussed in the previous subsection, if $y_k^*(t)$ denotes the emulated value of $y_k(t)$ it is given by equation

(2.2.7) in the Laplace domain. The term $E_k B/C$ in equation (2.2.7) is not a proper transfer function for $k > \rho$. This term can be decomposed, using polynomial long division, into two parts

$$\frac{E_k B}{C} = H_k + \frac{G_k}{C} \quad (2.3.3)$$

where G_k/C is strictly proper, H_k is the remainder polynomial and

$$\begin{aligned} \deg(H_k) &= k - \rho \\ \deg(G_k) &= n - 2 \\ n &\neq \deg(A) \end{aligned}$$

The emulator equation (2.2.7) becomes

$$Y_k^*(s) = H_k U(s) + \frac{G_k}{C} U(s) + \frac{E_k}{C} Y(s) \quad (2.3.4)$$

Notice that the emulator equation has two parts: one part can be realized by using proper transfer functions, the other part cannot. Using this fact, equation (2.3.4) can be rewritten as

$$Y_k^*(s) = H_k U(s) + Y_k^0(s) \quad (2.3.5)$$

where

$$Y_k^0(s) = \frac{G_k}{C} U(s) + \frac{E_k}{C} Y(s) \quad (2.3.6)$$

is the realizable part.

In the time domain, the emulator equation (2.3.5) becomes

$$y_k^*(t) = \mathbf{h}_k \mathbf{u} + y_k^0(t) \quad (2.3.7)$$

where \mathbf{h}_k is a row vector and contains the coefficients of the H_k polynomial and \mathbf{u} is a column vector which contains the input derivatives. Indeed, the elements of \mathbf{h}_k are the system Markov parameters

$$\mathbf{u} = [u(t) \dot{u}(t) \ddot{u}(t) \dots u^{(N_p)}(t)]^T, \quad u_k(t) = \frac{d^k u(t)}{dt^k} \quad (2.3.8)$$

By substituting equation (2.3.7) into equation (2.3.1), and rearranging in a matrix form, the following expression for the T -ahead predictor arises

$$\mathbf{y}^*(t+T) = \mathbf{T}_{N_p} H \mathbf{u} + \mathbf{T}_{N_p} \mathbf{Y}^0 \quad (2.3.9)$$

where

$$\mathbf{T}_{N_p} = \begin{bmatrix} 1 & T & \frac{T^2}{2!} & \dots & \frac{T^{N_p}}{N_p!} \end{bmatrix} \quad (2.3.10)$$

$$\mathbf{Y}^0 = [y(t) y_1^0(t) \dots y_{N_p}^0(t)]^T \quad (2.3.11)$$

and H is the $(N_p + 1) \times (N_p - \rho + 1)$ coefficients matrix of the polynomials H_k . When $\rho = 1$, the

[†] A function $f(t)$ can be approximated about a specified point t_0 by its derivatives at that point. This is known as the Taylor series expansion of that function about that point. If $t_0 = 0$ the series is called the Maclaurin series. In our case, the derivatives of the output $y(t)$ (at time t) are known. So we are defining a new time axis T (taking the time t as the origin) and doing the expansion in this new variable T .

matrix H is given as follows

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ h_1 & 0 & 0 & 0 & \cdots & 0 \\ h_2 & h_1 & 0 & 0 & \cdots & 0 \\ h_3 & h_2 & h_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{N_v} & \vdots & \vdots & \vdots & \cdots & h_1 \end{bmatrix} \quad (2.3.12)$$

where the h_k s are the Markov parameters of the open-loop system B/A . This can easily be proven by using the identities (2.2.2) and (2.3.3). If $\rho > 1$ then $h_k = 0$ for $k < \rho$ and as a result the number of the columns of the matrix H decreases with ρ .

Equation (2.3.9) is the basic equation in the development of CGPC control law, however the following Laplace transform form of it will be needed to write the resulting time domain CGPC control law in the Laplace domain and thus to develop the closed-loop equations.

$$Y_T^*(s) = T_{N_v} H S_H U(s) + T_{N_v} Y^0(s) \quad (2.3.13)$$

where

$$S_H = [1 \ s \ s^2 \ \cdots \ s^{N_v - n}]^T \quad (2.3.14)$$

$$Y^0(s) = \frac{G S_G}{C(s)} U(s) + \frac{F S_f}{C(s)} Y(s) \quad (2.3.15)$$

where G and F are $(N_v + 1) \times (n - 1)$, $(N_v + 1) \times n$ coefficient matrices of G_k and F_k polynomials respectively and S_G , S_f are corresponding s vectors.

$$S_G = [s^{n-2} \ s^{n-3} \ \cdots \ 1]^T \quad (2.3.16)$$

$$S_f = [s^{n-1} \ s^{n-2} \ \cdots \ 1]^T. \quad (2.3.17)$$

The subscript T is used in equation 2.3.13 to indicate that the Laplace transform is taken with respect to t and T is left as a parameter.

Control order. In the discrete-time GPC (Clarke *et al.*, 1987), the *control horizon* (N_u) is a key design parameter. The significance of this parameter is that the predicted control† increments are constrained to be zero (Clarke *et al.*, 1987) for future time instants greater than (N_u). This constraint is convenient for the

following reasons:

1. It reduces the dimension of the matrix involved in the control law calculations and thus the computational burden.
2. It enables the control of non-minimum phase systems without other forms of control weighting.
3. It can be used to adjust the system transients.

As discussed by Clarke *et al.* (1987), more active control action results from larger N_u and vice versa. As a result, N_u is a useful design parameter.

In discrete-time, the predictor equation explicitly includes the future controls whereas in continuous-time (equation 2.3.9) the future control is implicitly included in terms of current input derivatives and the future variable T . More precisely, the future control (predicted control) appears to be a *polynomial* in T . Therefore the above discrete-time constraint is not as appropriate for the continuous-time case. Instead, we use the constraint that input *derivatives* of order greater than N_u are zero that is

$$u_k(t) = 0 \quad \text{for } k > N_u. \quad (2.3.18)$$

We call this N_u the “control order” because of the obvious reason that the predicted control is constrained to be a polynomial of order N_u . For example, the predicted control will be a constant for $N_u = 0$, a ramp for $N_u = 1$ and so on.

The control order (N_u) is *algebraically* equivalent to the control horizon (N_u) because it reduces the dimension of the vector u to $(N_u + 1) \times 1$ and the dimension of the matrix H to $(N_v + 1) \times (N_u + 1)$. However, they are not *physically* equivalent; despite this, the constraint (2.3.18) has similar effects in continuous-time (control of non-minimum phase systems with zero control weighting, the larger N_u the faster response etc.). This will be discussed in detail later in the paper.

Interpretation of the predictor. At any time t , the future response of any linear system can be divided into three parts.

$$y(t + T) = y_u(t, T) + y_i(t, T) + v(t + T) \quad (2.3.19)$$

where:

- $y_u(t, T)$ is the response to the input after time t assuming zero initial conditions at time t
- $y_i(t, T)$ is the response to initial conditions at time t created by the past data, assuming zero input after time t
- $v(t + T)$ is the future noise component.

At time t , $y_i(t, T)$ is exactly known, $y_u(t, T)$ depends on the future input and $v(t + T)$ is not

† The term “predicted control” is used throughout the paper for the future control which is to be calculated, at time t , from the predictor model in order to minimise the specified cost function. The calculations are based on the receding horizon strategy: the predicted control $u^*(t, T)$ is *not* applied to the system over the time interval for which the cost function is minimized, but rather only its value at $T = 0$ ($u^*(t, 0)$) is applied. Therefore the predicted control and the actual future control are not the same. This matter will be clarified in Section 2.4.

known. Thus, assuming that future noise is zero, a predictor model for the system depending on the future control is obtained as

$$\hat{y}(t+T) = h(T) * u(t+T) + y(t, T) \quad (2.3.20)$$

where $h(T)$ is the impulse response of the system and $*$ denotes convolution.

The predictor of equation (2.3.9) is of exactly the same form. The term $T_N Y^0$ entirely depends on the past data and can be calculated at time t ; thus this part is related to the initial condition response. The part $T_N H u$ depends on the future input in terms of input derivatives at time t and the future variable T . Thus it is an approximation to $h(T) * u(t+T)$. To make this point more clearly, consider the following approximate functions.

$$h(T) \approx \tilde{h}(T) = h_1 + h_2 T + h_3 \frac{T^2}{2!} + \dots + h_{N_s} \frac{T^{(N_s-1)}}{(N_s-1)!} \quad (2.3.21)$$

$$u(t+T) \approx \tilde{u}(t+T) = u(t) + u_1(t)T + u_2(t) \frac{T^2}{2!} + \dots + u_{N_u}(t) \frac{T^{N_u}}{N_u!} \quad (2.3.22)$$

It can be shown that

$$\tilde{h}(T) * \tilde{u}(t+T) = T_N H u + T_r H_r u \quad (2.3.23)$$

T_r is given by

$$T_r = \left[\frac{T^{(N_s+1)}}{(N_s+1)!}, \dots, \frac{T^{(N_s+N_u)}}{(N_s+N_u)!} \right] \quad (2.3.24)$$

and H_r by

$$H_r = \begin{bmatrix} 0 & h_{N_s} & h_{(N_s-1)} & \dots & h_{(N_s-N_u+1)} \\ 0 & 0 & h_{N_s} & \dots & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & h_{(N_s-1)} \\ 0 & 0 & 0 & \dots & h_{N_s} \end{bmatrix} \quad (2.3.25)$$

Note that H is of dimension $(N_s+1) \times (N_u+1)$ as discussed in the previous subsection.

If $N_s > N_u$ (which is the case in the CGPC), then the term $T_r H_r u$ will be much smaller than the term $T_N H u$. In particular, when $N_s \rightarrow \infty$, $T_r H_r u \rightarrow 0$. So, by ignoring this term one can obtain the following approximation:

$$y_u(t, Y) = h(T) * u(t+T) = T_N H u \quad (2.3.26)$$

The accuracy of this approximation depends on the parameters N_s , N_u and the range of T . It is

worth emphasising that the above approximation can be made very close by the proper choice of these parameters.

2.4. Reference trajectory

The objective of the CGPC, as in the discrete-time GPC, is to drive the predicted future output as close as possible to the future setpoint subject to the input constraints. This implies that the future setpoint needs to be known, which is the case in some applications such as robotics. However, in many applications, the future setpoint is not known. In this case, one could consider a constant setpoint w into the future, but trying to match the predicted output to a constant value might give an excessive control action or overshoot at the output. The alternative approach used here is to consider a reference output which goes smoothly from the current output $y(t)$ to w as illustrated in Fig. 1. As will be seen later, this approach indeed has the effect of reducing the overshoot and the control activity, in addition it enables us to obtain model-following type control (even sometimes exact model-following) with the right choice of CGPC design parameters.

The reference trajectory $w_r(t, T)$ will be taken as the output of a rational transfer function (reference model) with numerator R_n and denominator R_d :

$$W_r(t, s) = \frac{R_n(s) w(t) + y(t)}{R_d(s) s} \quad (2.4.1)$$

Note that here the Laplace operator s denotes the Laplace transform with respect to future variable T , not t . In order to have the same structure as the output predictor (equation 2.3.9), the reference trajectory will be approximated as a truncated Maclaurin series. Consider the following approximation to R_n/R_d

$$\frac{R_n(s)}{R_d(s)} \approx \sum_{i=0}^{N_r} r_i s^i \quad (2.4.2)$$

where r_i are the Markov parameters of R_n/R_d .

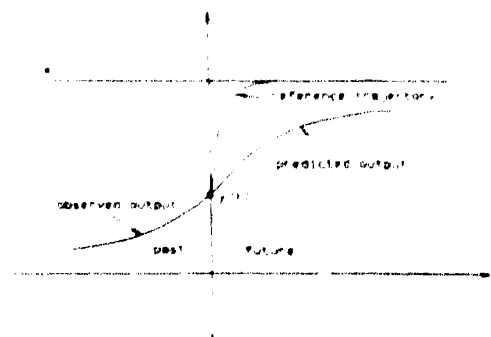


Fig. 1. Graphical illustration of CGPC strategy

By substituting equation (2.4.2) into equation (2.4.1) and taking the inverse Laplace transform with respect to T (and rearranging in a matrix form) the following approximation to the reference trajectory is found

$$w_r^*(t, T) = \mathbf{T}_N \mathbf{w} \quad (2.4.3)$$

where

\mathbf{T}_N is a row vector as defined before

$$\text{and} \quad \mathbf{w} = \mathbf{r}[w(t) - y(t)] \quad (2.4.4)$$

where \mathbf{r} is a column vector which contains the Markov parameters of $R_n(s)/R_d(s)$.

$$\mathbf{r} = [r_0 \ r_1 \ \dots \ r_N]^T. \quad (2.4.5)$$

If the future setpoint is known, then this can be used instead of the reference output. To do this, the future setpoint should be approximated as a truncated Maclaurin series as in the form of equation (2.4.3). This can then be used in the control law calculations. However, if the future setpoint does not have continuous derivatives in the given time frame, this approximation is not possible; unless the problem is divided into intervals in which the future setpoint has continuous derivatives. This implies that the cost function given in the next section should also be divided accordingly. Therefore, setpoints which are discontinuous or have discontinuous derivatives, such as a square wave, cannot easily be incorporated into the design, unlike discrete-time GPC.

Remark. In some control methods a setpoint prefilter is used to adjust the closed-loop setpoint response. In the GMV method the setpoint weighting transfer function R also comes into the controller in this way. This is not so in the CGPC: the effect of the transfer function R_n/R_d implicitly appears in the controller gain. This point will become clear after the control law calculations.

2.5. Control law

The CGPC, like the GPC, is based on a receding time frame. That is at a given time t the cost function minimization occurs not with respect to t but with respect to a receding time frame whose origin is at time t . Thus for each time t , a pseudo input $u_r(t, T)$, a pseudo output $y_r(t, T)$ and a pseudo setpoint $w_r(t, T)$ are considered where T is the receding time variable and t is a constant for that time frame. These pseudo variables are defined so as to be directly related to the actual system variables at $T = 0$ (see equation 2.5.4 and Fig. 1.):

$$y_r(t, 0) = 0 \quad (2.5.1)$$

$$u_r(t, 0) = u(t). \quad (2.5.2)$$

These pseudo variables are undefined for $T < 0$ and have no direct relationship with the actual variables for $T > 0$. In particular, it is *not* generally true that $u_r(t, T) = u(t + T)$.

In the discrete-time GPC, output prediction† depends on the future controls which are to be determined. Suppose that the future controls are known, then the predicted output can be calculated. The reverse operation is also possible: given a predicted output over a time frame, the corresponding future controls can be calculated. These future controls will be called predicted controls. GPC does this reverse operation by minimising a cost function over the given time frame. The first element $u(t)$ of the predicted controls is then applied to the system and the same procedure is repeated at the next time instant.

Similarly, in continuous-time, the predicted output depends on the input $u(t)$ and its derivatives (see the predictor equation 2.3.9). In other words, future control (predicted control) is a polynomial of order N_u in T . If the input and its derivatives are known, the predicted output can be calculated; or given the output prediction over a time frame, the corresponding input and input derivatives can be calculated. The objective of the CGPC is then to find the input and its derivatives such that predicted output is as close as possible to the reference trajectory. This is done by minimizing a cost function (equation 2.5.3), similar to the one in discrete case, over the given time frame. Having obtained the Maclaurin representation of the pseudo control, only the first term of the series is used in computing $u(t)$ from 2.5.2.

Finding the input and its derivatives is equivalent to finding the predicted control $u_r^*(t, T)$. Because the predicted control is in the form of a truncated Maclaurin series of order N_u , the CGPC control input can be written as $u(t) = u_r^*(t, 0)$.

Consider the following cost function

$$J = \int_{t_1}^{t_2} [y_r^*(t, T) - w_r^*(t, T)]^2 dT + \lambda \int_0^{t_2 - t_1} [u_r^*(t, T)]^2 dT \quad (2.5.3)$$

where

$$y_r^*(t, T) = y^*(t + T) - y(t) \quad (2.5.4)$$

$$u_r^*(t, T) = \sum_{k=0}^{N_u} u_k(t) \frac{T^k}{k!} \quad (2.5.5)$$

† The term "output prediction" is used to imply the j -step ahead (T -ahead for the continuous-time case) output prediction based on the information available at time t , where $j(T)$ is any integer (real) number greater than the system time delay.

or in matrix form

$$u_r^*(t, T) = \mathbf{T}_{N_r} \mathbf{u} \quad (2.5.6)$$

$$\mathbf{T}_{N_r} = \begin{bmatrix} 1 & T & \frac{T^2}{2!} & \cdots & \frac{T^{N_r}}{N_r!} \end{bmatrix} \quad (2.5.7)$$

$$\mathbf{u} = [u(t) \ u_1(t) \ \cdots \ u_{N_r}(t)]^T \quad (2.5.8)$$

T_1 = minimum prediction horizon

T_2 = maximum prediction horizon

λ = control weighting.

Note that $y_r^*(t, T)$ is exactly given by the same predictor equation (equation 2.3.9) except that the first element of \mathbf{Y}^0 is set to zero. For the Laplace domain equation (equation 2.3.13) this is equivalent to setting the first rows of G and F matrices to zero.

Remarks.

1. Note that a minimum prediction horizon T_1 is included in the cost. T_1 can always be chosen to be zero; but, if the system is known to have a delay, T_1 can be set equal to that delay. If the time delay is not known then T_1 can be set equal to the *largest* possible delay.
2. The situation in Remark 1 is rather different from that found in the discrete-time case where, if the time delay is not known, N_1 is set to the minimal delay. This is because if the actual delay is large the corresponding leading elements of B are estimated as zero. If N_1 set to the maximum delay and if the actual delay is smaller then the B polynomial will be underparameterised and this may create estimation problems. The source of this difference is that in discrete time delay corresponds to relative order. In continuous-time the same problem occurs if the order of $B(s)$ is assumed to be smaller than the actual order of B , however the problem does *not* occur in the time delay case. Hence, we believe that T_1 should be set to the maximum possible delay to exclude possible range of time delay variations.
3. Returning to the continuous-time case, the predicted input $u_r^*(t, T)$ in the interval $T > T_2 - T_1$ has no effect on the predicted output $y_r^*(t, T)$ within the range covered by the cost function as long as T_1 is equal to the system time delay. This is the reason for the choice of the integration ranges in the cost function 2.5.3.
4. In the cost the predicted output (not the output itself) is considered. In this way, we transfer a stochastic problem into a deterministic framework. However, this does not mean that the effect of disturbances is ignored: the predicted output takes into

account the effect of disturbances up to time t and thus the effect of disturbances are included in the cost indirectly.

The CGPC control law can now be restated as follows.

1. Find the vector \mathbf{u} which minimises the above cost (2.4.3)
2. Use the first element of \mathbf{u} $u(t)$ as control input.

With the substitution of equations (2.3.9), (2.4.3) and (2.5.6) into equation (2.5.3), the cost becomes

$$J = \int_{T_1}^{T_2} (\mathbf{T}_{N_r} H \mathbf{u} + \mathbf{T}_{N_r} \mathbf{Y}^0 - \mathbf{T}_{N_r} \mathbf{w})^T dT + \lambda \int_{T_1}^{T_2 - T_1} \mathbf{u}^T \mathbf{T}_{N_u}^T \mathbf{T}_{N_u} \mathbf{u} dT \quad (2.5.9)$$

The minimization of the J results in

$$\mathbf{u} = \mathbf{K}(\mathbf{w} - \mathbf{Y}^0) \quad (2.5.10)$$

where

$$\mathbf{K} = (H^T \mathbf{T}_1 H + \lambda \mathbf{T}_{N_u})^{-1} H^T \mathbf{T}_1 \quad (2.5.11)$$

$$\mathbf{T}_1 = \int_{T_1}^T \mathbf{T}_{N_r}^T \mathbf{T}_{N_r} dT \text{ is } (N_r + 1) \times (N_r + 1) \quad (2.5.12)$$

$$\mathbf{T}_{N_u} = \int_{T_1}^{T_2 - T_1} \mathbf{T}_{N_u}^T \mathbf{T}_{N_u} dT \text{ is } (N_u + 1) \times (N_u + 1). \quad (2.5.13)$$

Note that \mathbf{T}_{N_u} becomes submatrix of \mathbf{T}_1 when $T_1 = 0$. Let the first row of \mathbf{K} be \mathbf{k} , then CGPC control law is given by

$$u(t) = \mathbf{k}[\mathbf{w} - \mathbf{Y}^0]. \quad (2.5.14)$$

In the Laplace domain

$$U(s) = \mathbf{k}r[W(s) - Y(s)] - \mathbf{kY}^0(s) \quad (2.5.15)$$

By substituting equation (2.3.15) into equation (2.5.15), the following transfer function form of the CGPC control law is obtained

$$U(s) = g[W(s) - Y(s)] - \frac{G_0}{C} U(s) - \frac{F_0}{C} Y(s) \quad (2.5.16)$$

where the scalar gain g and the polynomials G_0 and F_0 are given by

$$g = \mathbf{k}r \quad (2.5.17)$$

$$G_0 = \mathbf{kGS}_r \quad (2.5.18)$$

$$F_0 = \mathbf{kFS}_r \quad (2.5.19)$$

The feedback system given by the CGPC control law (equation 2.5.16) is illustrated in Fig. 2. The CGPC control law can also be rewritten in the following form

$$W(s) = \Phi_r(s) \quad (2.5.20)$$

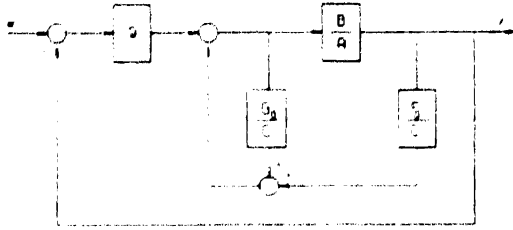


FIG. 2. The feedback system of CGPC

where

$$\Phi_r(s) = \frac{G_r}{C} U(s) + \frac{F_r}{C} Y(s) \quad (2.5.21)$$

$$G_r = \frac{G_0 + C}{g} \quad (2.5.22)$$

$$F_r = \frac{F_0}{g} + C. \quad (2.5.23)$$

Remarks.

1. Note that the CGPC control law is obtained by setting the output of an equivalent emulator equal to the setpoint (equation 2.5.20). The equivalent emulator polynomials satisfy the pole-placement identity $PC = G_r A + F_r B$ but P is implicitly specified by the CGPC algorithm.
2. An important result of using the receding time frame is that, although the control law required to realize $u_r^*(t, T)$ (for fixed t and varying T) would vary with T , the actual control law required to realize $u(t) = u_r^*(t, 0)$ for variable t and fixed $T = 0$ does not depend on t . Thus, like the GPC, the CGPC is a time invariant control law.
3. Note that the effect of R_n/R_d is reflected by the vector \mathbf{r} in the gain \mathbf{g} .

3. ANALYSIS OF THE CGPC CLOSED-LOOP SYSTEM

3.1. General closed-loop equations

Closed-loop setpoint response. The CGPC control law gives the following closed-loop response to the setpoint.

$$Y(s) = \frac{gBC}{AC + G_0A + F_0B + gBC} W(s). \quad (3.1.1)$$

Using the decomposition identities (equation 2.2.2 and 2.3.3), one can show the following relationship

$$G_k A + F_k B = L_k C \quad (3.1.2)$$

where L_k satisfies

$$\frac{s^k B}{A} = H_k + \frac{L_k}{A}. \quad (3.1.3)$$

Considering equation (3.1.2), together with the equations for the polynomials G_0 and F_0 (equations 2.5.18 and 2.5.19), it follows that

$$G_0 A + F_0 B = L_0 C \quad (3.1.4)$$

where the polynomial L_0 is given by

$$L_0 = \mathbf{k} L S_L. \quad (3.1.5)$$

L is the $(N_y + 1) \times n$ coefficient matrix of the polynomials L_k ($k = 1, \dots, N_y$ and the first row of L is zero) and S_L is the corresponding $n \times 1$ vector containing powers of s .

$$S_L = [s^{n-1} s^{n-2} \dots s^1]^T. \quad (3.1.6)$$

The order of L_0 is $n-1$. It follows from equation (3.1.4) that C is a factor of both numerator and denominator of the closed-loop system. Cancellation of this common factor results in the following closed-loop transfer function:

$$Y(s) = \frac{gB}{A + L_0 + gB} W(s). \quad (3.1.7)$$

The feedback system given by equation (3.1.7) is shown in Fig. 3.

Note that the closed-loop zeros are equal to the open-loop zeros (except for one special case which will be given in the next subsection). In addition to this, the closed-loop system has the same degree as the open-loop system. State feedback would give the same effect. Indeed, the feedback configuration given in Fig. 3 corresponds to a state feedback where the partial state and its derivatives (Wolovich, 1974; Kailath, 1980) are fed back through the gain vector $\mathbf{k}(L + \mathbf{r}B)$ where B is the $1 \times n$ row vector of the coefficients of the polynomial B . Hence, the CGPC control law is equivalent to state feedback for a particular set of state feedback gains. In essence, this state feedback is made possible by the polynomial $C(s)$, which acts as an observer polynomial yielding state information from the input-output data.

Using identity (3.1.3), the polynomial L_0 can be written as

$$L_0 = B\mathbf{k}S - A\mathbf{k}H_r S_{H_r} \quad (3.1.8)$$

where

$$S = [0 \ s \ s^2 \ \dots \ s^{N_y}]^T. \quad (3.1.9)$$

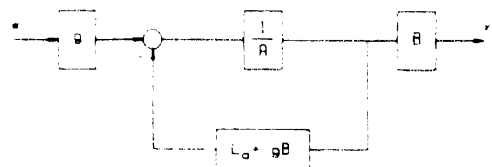


FIG. 3. Equivalent CGPC feedback system.

H_f is the full H matrix; that is the H matrix when $N_u = N_s - \rho$ and S_{H_f} is the corresponding s vector. In addition, L_0 can be written as

$$L_0 = BZ_0 - AH_0 \quad (3.1.10)$$

where the polynomials Z_0 and H_0 are defined as follows

$$Z_0 = \mathbf{k}S \quad (3.1.11)$$

$$H_0 = \mathbf{k}H_f S_{H_f} \quad (3.1.12)$$

With the substitution of equation (3.1.10) into equation (3.1.7), a new expression for the closed-loop system is obtained as follows

$$Y(s) = \frac{gB}{A(1 - H_0) + B(Z_0 + g)} W(s). \quad (3.1.13)$$

This expression is used to prove some properties of the closed-loop system in the following subsections.

Closed-loop disturbance response. The closed-loop disturbance response is given by

$$Y(s) = \frac{(G_0 + C)A}{(A + L_0 + gB)C} \Psi(s) \quad (3.1.14)$$

where $\Psi(s)$ is the direct disturbance at the output and given by $\Psi(s) = (C/A)V(s)$. Equation (3.1.14) can be divided into two parts

$$Y(s) = \frac{A}{A + L_0 + gB} \Psi(s) + \frac{AG_0}{(A + L_0 + gB)C} \Psi(s) \quad (3.1.15)$$

Although the second part of the closed-loop disturbance response may be adjusted by C polynomial without affecting the closed-loop setpoint response, as a whole we do not have enough flexibility to adjust the disturbance transients separately from the closed-loop setpoint response.

Closed-loop control input. The closed-loop control input is given by

$$U(s) = \frac{gA}{A + L_0 + gB} W(s) - \frac{gC + F_0}{A + L_0 + gB} V(s). \quad (3.1.16)$$

3.2. A special case

A special case occurs when $\lambda = 0$ and $N_u = N_s - \rho$: the CGPC control law becomes a zero cancellation law: the closed-loop pole polynomial has B as a factor and would give unstable control for non-minimum phase systems. In this special case, the closed-loop system is given as follows.

$$Y(s) = \frac{1}{Z(s)} W(s) \quad (3.2.1)$$

where

$$Z(s) = \frac{Z_0(s) + g}{g} \quad (3.2.2)$$

This may be demonstrated as follows.

Consider the gain matrix K (equation 2.5.11) when $\lambda = 0$ and $N_u = N_s - \rho$

$$K = (H_f^T T_1 H_f)^{-1} H_f^T T_1 \quad (3.2.3)$$

It follows from equation (3.2.3) and (3.1.12) that the polynomial $H_0 = 1$. This completes the proof (see equation 3.1.13). Note that $\deg(Z(s)) = \rho$.

In general, the analysis of the closed-loop pole locations in terms of the CGPC design parameters seems impossible analytically. In this special case, however, this can be done to some extent. This may also provide some insight into the general case.

Consider the matrix H_f ; it can be decomposed as follows

$$H_f = \begin{bmatrix} H_{f1} \\ H_{f2} \end{bmatrix} \quad (3.2.4)$$

where H_{f1} is a zero matrix with the dimension $\rho \times (N_s - \rho + 1)$, and H_{f2} is a lower triangular square matrix in the following form

$$H_{f2} = \begin{bmatrix} h_{\rho} & 0 & \cdots & 0 \\ h_{\rho+1} & h_{\rho} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{N_s} & \cdots & \cdots & h_{\rho} \end{bmatrix} \quad (3.2.5)$$

The matrix T_1 can also be decomposed appropriately as

$$T_1 = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \quad (3.2.6)$$

It can be shown that the gain matrix K (equation 3.2.3) can be written as follows

$$K = H_{f2}^{-1} [T_{12}^{-1} T_{21} I] \quad (3.2.7)$$

where I is a unit matrix with appropriate dimension. Since H_{f2} is lower triangular its inverse is also lower triangular; hence the first row of K is given as

$$\mathbf{k} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} [T_{12}^{-1} T_{21} I] \quad (3.2.8)$$

or

$$\mathbf{k} = \frac{1}{h_{\rho}} [\mathbf{T}_1 \ 1 \ 0 \ \cdots \ 0]. \quad (3.2.9)$$

\mathbf{T}_1 is the first row of $T_{12}^{-1} T_{21}$. Note that the dimension of \mathbf{T}_1 is $1 \times \rho$.

The elements of \mathbf{T}_1 are non-linear functions of different powers of T_1 and T_2 , so even in this simple case a general expression for the

closed-loop pole polynomial (in this case $Z(s)$) will be quite complex. Therefore we will not examine $Z(s)$ in the general case; instead we will consider two special cases where $\rho = 1$ and $\rho = 2$. In both cases R_n/R_d is assumed to be a first order transfer function.

$$\frac{R_n(s)}{R_d(s)} = \frac{1}{rs + 1} \quad (3.2.10)$$

One can show that, for this R_n/R_d , the vector r will be

$$0 \frac{1}{r} \frac{-1}{r^2} \frac{1}{r^3} \dots \frac{-1}{(-r)^N} \quad (3.2.11)$$

Case 1. Here $\rho = 1$ and thus T_1 is scalar, say t_1 . Then it follows from equation (2.5.17) and (3.1.11) that

$$g = \frac{1}{h_\rho r}; \quad Z_0 = \frac{s}{h_\rho} \quad (3.2.12)$$

Hence, from equation (3.2.2)

$$Z(s) = rs + 1. \quad (3.2.13)$$

This means that under these circumstances we obtain *exact model-following* regardless to the choice of T_1 and T_2 . Note that if $R_n/R_d = 1$ then $Z(s)$ will be

$$Z(s) = \frac{1}{t_1} s + 1 \quad (3.2.14)$$

and the pole location can be adjusted by T_1 and T_2 : generally T_1 is chosen to be zero and T_2 is used for adjustment. Simulation results show that the effect of T_2 on the pole location is that the smaller the T_2 , the further away the pole from the imaginary axis.

Case 2. Here $\rho = 2$, thus $T_1 = [t_{11} \ t_{12}]$, hence

$$g = \frac{1}{h_\rho} \left(\frac{t_{12}}{r} - \frac{1}{r^2} \right); \quad Z_0 = \frac{1}{h_\rho} (s^2 + t_{12}s). \quad (3.2.15)$$

One can then show that, in this case, $Z(s)$ is given as follows

$$Z(s) = (rs + 1) \left(\frac{r}{rt_{12} - 1} s + 1 \right). \quad (3.2.16)$$

Equation (3.2.16) shows that, in this case, irrespective of T_1 and T_2 , one of the closed-loop poles will be at the location defined by R_d . The other pole can be placed far away from the imaginary axis by the proper choice of T_1 and T_2 . This results in very close model-following. This is also true for $\rho > 2$, that is one of the closed-loop poles will be at the location defined by R_d and, replacing the other poles further left from this model-following type control can be obtained to some extent. Note that if $R_n/R_d = 1$

then $Z(s)$ is given by

$$Z(s) = \frac{1}{t_{11}} (s^2 + t_{12}s + t_{11}). \quad (3.2.17)$$

3.3. The effect of common factors

If the system model is overspecified, a common factor will appear in the estimated model when self-tuning control is used. It is therefore important to examine the case where the system *does not* have a common factor, but the model on which the design is based *does* have a common factor. Consider the following model

$$A = A'X; \quad B = B'X \quad (3.3.1)$$

where A' and B' are the actual system polynomials and X is a common factor. There are two questions to be answered:

1. Does the common factor create any problem in the control law calculations?
2. How does the common factor affect the control law?

Examination of the decomposition identities (equations 2.2.2 and 2.3.3) shows that common factors will not create any problem in the solution but, for different common factors, we will have different F_k and G_k polynomials. Although H_k polynomials and thus the vector k do not depend on the common factor [this is apparent from the identity (3.1.3)], this gives rise to different G_0 and F_0 polynomials for different common factors. Note that gain g is independent of common factor.

As the control law is applied to the actual system, the closed-loop system (equation 3.1.1) will be given in terms of the actual system polynomials A' and B' . So we will have the term $G_0A' + F_0B'$ in the denominator instead of $G_0A + F_0B$. What this term will be as G_0 and F_0 are different for different common factors? Examination of the identity (3.1.3) shows that

$$L_0 = L_0'X \quad (3.3.2)$$

where L_0' is the L_0 polynomial when $X = 1$. It then follows from equation (3.1.4) that

$$G_0A' + F_0B' = L_0'C. \quad (3.3.3)$$

Equation (3.3.3) shows that the closed loop system does *not* depend on the common factor.

In the above analysis, we did not make any distinction between the stable and unstable common factors: the above result is true for both cases. Also, the above result is true when the system itself has a common factor but, in this case, unstable common factors will result in unstable control.

An important feature of CGPC is that, unlike

pole-placement, it does not suffer from the ill effects of the pole-zero cancellation.

3.4. Relation of CGPC to LQ control

CGPC is clearly related to LQ control. Here we give a brief heuristic discussion of this point. In Section 2.2 under the heading "Interpretation of the predictor", we discussed the relationship between the predicted future output $y^*(t+T)$ and the actual future output $y(t+T)$ and it was shown that $y^*(t+T)$ is an approximation to the noise-free future output $\hat{y}(t+T)$ (equation 2.3.20). It was also noted that the approximation accuracy depends on N_1 , N_u and T . From this argument, it is clear that when $N_1 \rightarrow \infty$ and $N_u = N_1 - \rho$ (largest possible N_u for a given N_1) the predicted output $y^*(t+T)$ can be replaced by $\hat{y}(t+T)$ (assuming no future noise) and in the same way the predicted control can be replaced by $u(t+T)$. By considering that t is the initial time and choosing $T_1 = 0$, then CGPC cost function can be written as†

$$= \int_0^{T_2} [y^2(T) + \lambda u^2(T)] dT \quad (3.4.1)$$

This is also LQ cost function for a single-input single-output system. The control law minimizing this cost function is a function of T . However, when $T_2 \rightarrow \infty$ the control law becomes stationary. Therefore, when $T_2 \rightarrow \infty$ and with the above choices of N_1 and N_u , CGPC and LQ control law becomes equivalent. This argument is also supported by simulation.

We can summarize the above discussion as. The following settings of the CGPC parameters results in LQ control.

$$R_u = R_d = 1$$

$$N_u = N_1 - \rho$$

$$N_1 \rightarrow \infty$$

$$T_1 = 0$$

$$T_2 \rightarrow \infty$$

4. THE EFFECTS AND CHOICE OF CGPC PARAMETERS

This section gives a practical guide to the choice of CGPC design parameters, and relates these parameters to the discrete-time GPC (Clarke *et al.*, 1987).

4.1. The minimum prediction horizon T_1

Usually, the minimum prediction horizon $T_1 = 0$; but it is useful to choose $T_1 > 0$ when the

system has a time delay or when it is non-minimum phase. If the system has a time delay then there is no point in setting T_1 less than the time delay since the corresponding output cannot be affected by $u(t)$. If the time delay is not known then T_1 may be chosen equal to the largest possible delay (Section 2.5). For non-minimum phase systems, T_1 may be chosen such that the negative going part is excluded. Although it is possible to obtain reasonable control for time delay and non-minimum phase systems with $T_1 = 0$, the above choice of $T_1 > 0$ will improve the control performance for each case. T_1 corresponds to N_1 in the discrete time formulation (Clarke *et al.*, 1987).

4.2. The maximum prediction horizon T_2

This parameter is equivalent to N_2 in the discrete-time formulation (Clarke *et al.*, 1987) and has the same effect in continuous-time. In general, the smaller value of T_2 corresponds to the faster output response and thus the more active control action. For the larger T_2 , the slower output response is obtained. Therefore, T_2 can be used as a knob to adjust the rise time of the closed-loop output response. However, if a reference trajectory is specified by R_u/R_d , the choice of T_2 needs to agree with the rise time of the model. If the system has an initially negative-going non-minimum phase response, the minimum value of T_2 should cover the later positive-going part.

4.3. The predictor order N_1

In the predictor design, the future output is approximated by a N_1^{th} order truncated Maclaurin series. It is obvious that approximation accuracy depends on N_1 . So N_1 needs to be chosen such that a good approximation can be obtained over the range in which T varies. In other words, the sum of the terms $y_i(t)(T_i^i/t^i)$ for $i = 0, \dots, N_1$ should be reasonably small. However, as $y_i(t)$ are not known, it seems impossible to choose N_1 on the basis of the above argument.

Intuitively, one may argue that if N_1 is chosen such that we have a good approximation of the open-loop system step (or impulse) response over the range $0 < T < T_2$, then this will also result in a good approximation for the output predictor. This is because the predictor design is based on the open-loop system. This intuitive argument can be supported as follows. Consider

†The setpoint has been omitted for simplicity since closed-loop stability properties are independent of the closed-loop system inputs.

‡ Approximation of a function $f(t)$ by the Taylor series, about a specified point t_0 , will be accurate near to that point and inaccurate away from that point. Of course, the range in which the approximation is good depends on the order of Taylor series (in our case N_1).

the output predictor (equation 2.3.9) when $N_u = 0$

$$y^*(t + T) = y_i^*(T)u(t) + T_{N_y}Y^0 \tag{4.3.1}$$

where

$$y_i^*(T) = T_{N_y}H = h_1T + H_2\frac{T^2}{2!} + \cdots + h_{N_y}\frac{T^{N_y}}{N_y!} \tag{4.3.2}$$

As one may notice, $y_i^*(T)$ is the approximate step response (truncated Maclaurin series) of the open-loop system. We believe that if N_y is chosen such that $y_i^*(T)$ is approximated well over the range $0 < T < T_2$, $T_{N_y}Y^0$ will also be approximated well over the same range as it is the initial condition response of the same system.

As a conclusion, choosing N_y such that the error between the real and approximate step responses of the open-loop system over the range $0 < T < T_2$ is reasonably small will be a good criteria. This is also supported by simulation results.

Consider the following illustrative example.

$$\frac{A(s)}{B(s)} = \frac{-0.2s + 1}{s(s^2 + 1)} \tag{4.3.3}$$

The actual and approximate step responses of this system for various N_y over the range $0 < T < 5$ is given in Fig. 4. As can be seen from the figure, for a larger T_2 a larger N_y is needed and vice versa. For example, for $T_2 = 5$, N_y should be 12 whereas for $T_2 = 3$, $N_y = 6$ will be sufficient.

It follows from the above discussion that there is a close link between T_2 and N_y . Therefore these two parameters should always be considered together. However simulation studies showed that a large number of systems can be controlled reasonably well with the value of $N_y = 6$. For the simple plants this value can be reduced even further. But, for the complex

systems (at the same time open-loop unstable, non-minimum phase and higher order) N_y can be as large as 30 depending on the complexity. This is the case generally with the time delay systems since their approximation becomes non-minimum phase and higher order. For example for the double integrator with the unit time delay (e^{-1/s^2}) N_y needs to be chosen around 20 when $T_2 = 2$.

Note that we do not use the step response representation in the predictor design; we introduce the step response only to give a criteria in order to choose N_y . However, it is interesting to note that N_y will never be as large as in the control methods based on step or impulse response representation such as DMC or IDCOM (Cutler and Ramaker, 1980; Richalet *et al.*, 1978), since N_y is chosen in order to approximate a part of step (or impulse) response not all of it.

There is no *physical* equivalent to N_y in discrete time but *algebraically* N_y is equivalent to N_2 (maximum cost horizon) since both have the same effect on the dimension of the matrices.

4.4. The control order N_u

The control order N_u can be seen as a design parameter to constrain the predicted control $u_i^*(t, T)$. For example $N_u = 0$ constrains the predicted control (though not, of course, the actual control) to be constant in the future; $N_u = 1$ constrains the predicted control to be a ramp and so on. The smaller N_u , the greater the constraint on the predicted control and vice versa. Note that in this way we indirectly constrain the control $u(t)$. No mathematical argument has yet been devised, but it seems reasonable that more constraint on the predicted control $u_i^*(t)$ means more constraint on $u(t)$ or vice versa. As a result, a small value of N_u gives less active control $u(t)$ and slow output response. Increasing N_u makes the control and the corresponding output more active until a stage is reached where any further increase in N_u makes little difference. Simulation results show that a value of $N_u = 0$ gives generally acceptable control for a large variety of systems. But an increased value of N_u is needed for more complex systems (high order open-loop unstable and non-minimum phase systems.)

The control horizon N_u in discrete time is an alternative way of constraining the predicted control. Because of this, they have similar effects; but they are physically not the same. However, the alternatives are *mathematically* equivalent since both have the same effect on the dimension of matrices and thus on the control law calculations.

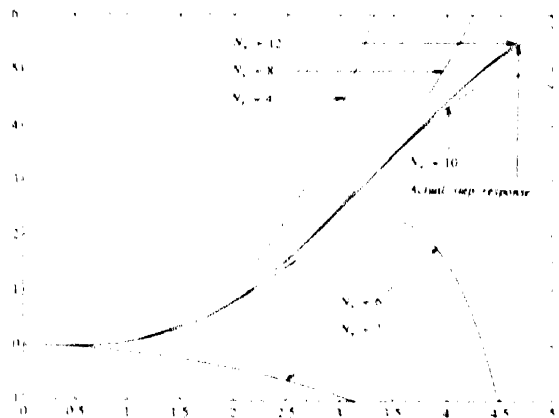


FIG. 4. Graphical illustration of the choice of N_y .

4.5. The model R_n/R_d

The closed-loop system response can be specified by a reference trajectory defined by a model R_n/R_d . The CGPC control law tries to match the system response to the model output. But, in general, it is not possible to obtain exact model-following since CGPC only changes the closed-loop pole locations. A special case where exact model-following is obtained is examined in Case 1 of Section 3.2. A first order R_n/R_d generally seems an appropriate choice. In this case, the CGPC places one of the closed-loop poles at the location specified by R_d , the rest away (the distance depends on the choice of T_2) from the imaginary axis when $N_u = N_k - \rho$ and $\lambda = 0$ (see Case 2 of Section 3.2). Thus it is possible to obtain a very close model-following with the right choice of T_2 . This model-following property becomes less accurate as N_u decreases from $N_k - \rho$. For a small N_u , such as 0 or 1, no longer model-following is obtained instead the effect of R_n/R_d is that it smooths the response by penalizing the overshoot.

5. A SIMULATION STUDY

5.1. Non-adaptive simulations

CGPC design parameters (N_k , N_u , I_1 , T_2 , λ , R_n/R_d) will have the same effects in both the adaptive and non-adaptive case. In this section a set of non-adaptive simulations are presented to illustrate the effects of these parameters on the properties of non-adaptive CGPC. Simulations for the self-tuning CGPC are given in Section 5.3. All of the simulations were performed using the MATLAB package program running on a Sun 3 workstation. The examples used in the adaptive and non-adaptive simulations are tabulated below.

Example 1:
$$\frac{B(s)}{A(s)} = \frac{1}{s(s^2 + 1)}$$
$$C(s) = 0.2s^2 + s + 1$$

Example 2:
$$\frac{B(s)}{A(s)} = \frac{-0.2s + 1}{s(s^2 + 1)}$$
$$C(s) = 0.2s^2 + s + 1$$

Example 3:
$$\frac{B(s)}{A(s)} = \frac{e^{-s}}{s^2}$$
$$C(s) = (s + 1)^3$$

Example 4:
$$\frac{B(s)}{A(s)} = \frac{1}{s^2 + 1}$$
$$C(s) = s + 1.$$

Note that the polynomial $C(s)$ is a design polynomial, not a part of the system.

5.1.1. The effects of T_2 and N_u . Example 1 was simulated to illustrate the effects of T_2 and N_u . In the simulations, the control weighting λ and T_1 were chosen to be zero, R_n/R_d to be 1 and the sample interval to be 0.1 sec.

Figure 5 illustrates the effect of T_2 where T_2 varies from 1 to 9 with an increment of 2. The predictor order N_k and the control order N_u were chosen to be 6 and 3 respectively, that is $N_k = 6$, $N_u = 3$. In the figure, the upper graph shows the step responses the lower graph shows the closed loop poles locus as T_2 varies. The fastest response in Fig. 5 corresponds to $T_2 = 1$. As can be seen from the graphs the response becomes slower and poles move towards the origin as T_2 increases.

If a model R_n/R_d is specified, the effect of T_2 may be different. This is because the specification given by the model and T_2 contradict with each other. In this case T_2 should be chosen by considering the time constant of the model.

The effect of N_u is shown in Fig. 6. For this example, N_k and T_2 were chosen to be 6 and 2 respectively and N_u is varied from 0 to 3. The upper graph shows the step responses and the lower one the poles locus as N_u varies. For

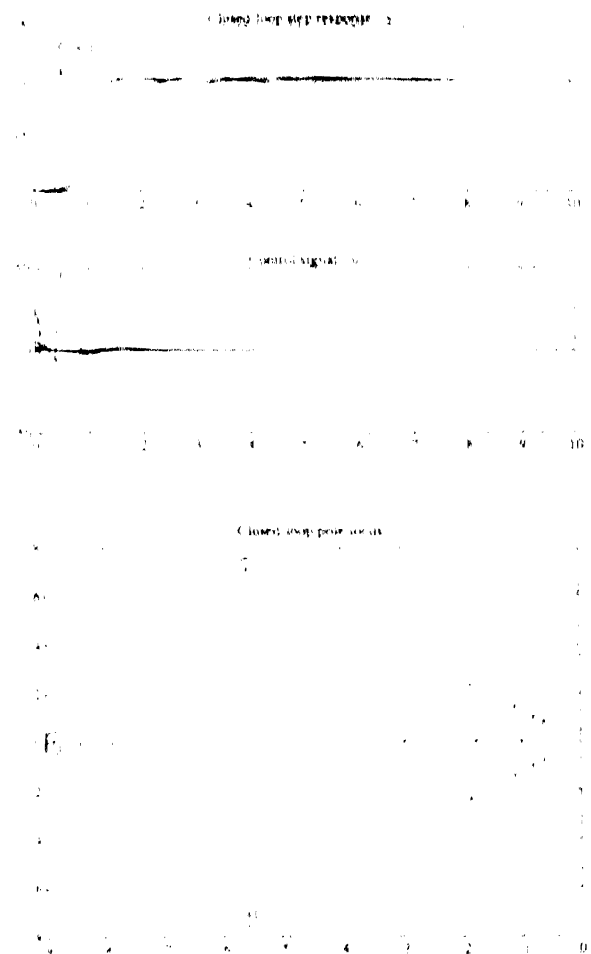


FIG. 5. Illustration of the effect of T_2 .

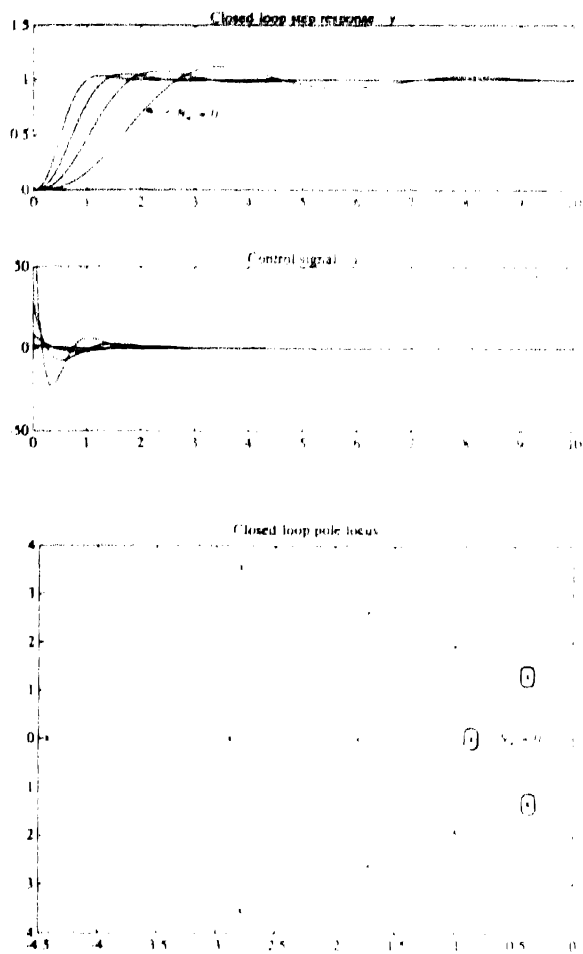


FIG. 6. The effect of the control order N_u .

$N_u = 0$, the output response and the corresponding poles are marked on the graphs. The output response become faster and the poles move away from the imaginary axis as N_u increases. Note that N_u and T_2 have opposite effects that is, when N_u is increased response becomes faster whereas when T_2 is increased response becomes slower.

5.1.2. *The effect of R_n/R_d .* As mentioned in Section 4.5, the reference model R_n/R_d has two functions: it can either be used as an approximate model or to penalise the overshoot. This section again uses Example 1 to illustrate these two functions of R_n/R_d . In the simulations, λ and T_1 were chosen to be 0, N_u to be 6, R_n to be 1 and the sample interval to be 0.1 sec.

Figure 7 shows the simulation result with $R_d = s + 1$ (model = $1/s + 1$), $T_2 = 1$ and $N_u = 3$. In the figure, poles are the corresponding closed-loop poles and ym is the model output. Note that CGPC control law placed one of the poles at -1 and the others far away from the imaginary axis in order to approximate the model. Also note the impulsive behaviour of the control at the initial time; this is because a third-order system was made to closely follow a

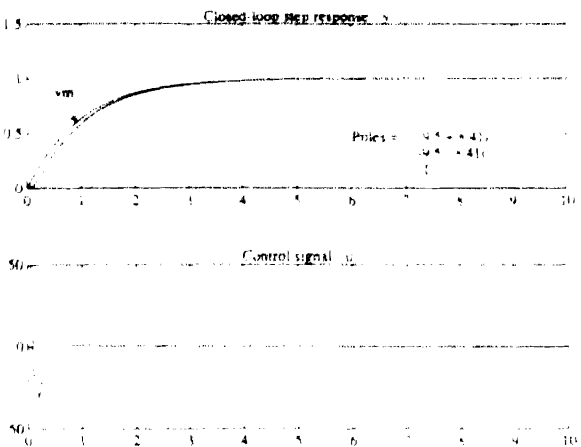


FIG. 7. The use of R_n/R_d as a reference model.

first-order model. In the example N_u was chosen to be $N_u = N_v - \rho$ in order to obtain the best approximation to the model. If N_u is chosen smaller this relationship becomes less accurate.

The use of R_n/R_d in order to penalise the overshoot is shown in Fig. 8. In this simulation, N_u and T_2 were chosen to be 0 and 1 respectively. The upper graph shows the step response when $R_d = 1$, the lower graph shows

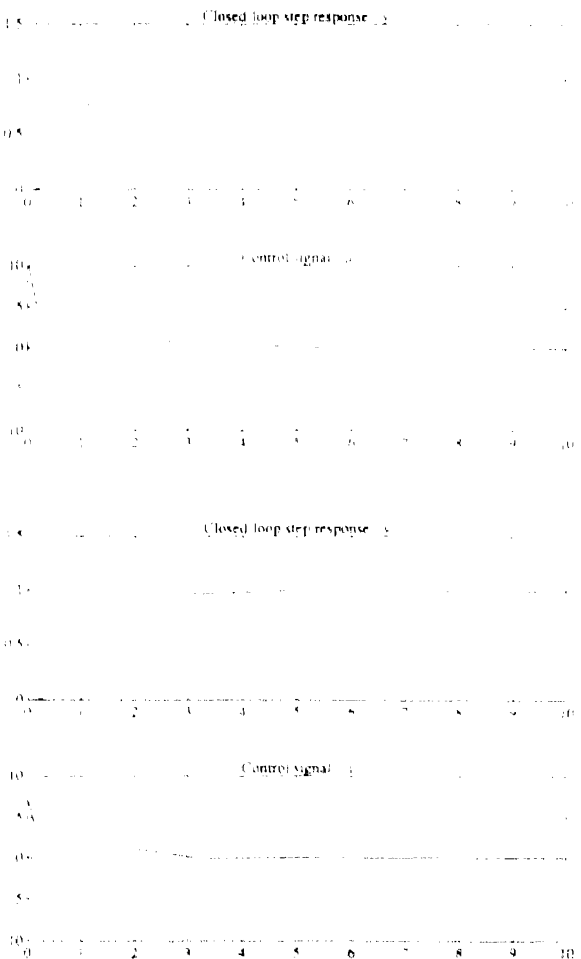


FIG. 8. The use of R_n/R_d to penalize the overshoot.

the step response when $R_d = 0.7s + 1$. Note that the use of R_d for this example, completely removed the overshoot.

5.1.3. Non-minimum phase systems. In contrast to the GMV algorithm, one of the properties of the GPC is that it is able to control non-minimum phase systems with zero control weighting. The CGPC also has this property. Example 2 was simulated to illustrate the control of non-minimum phase systems. In the simulation the following choice of parameters were used and the sample interval was chosen to be 0.05 sec.

$$\begin{aligned} R_u/R_d &= 1 \\ N_y &= 6 \\ N_u &= 0 \\ T_1 &= 0 \\ T_2 &= 3 \\ \lambda &= 0. \end{aligned}$$

The simulation result is illustrated in Fig. 9. As seen from the figure, output response is reasonable but it can further be improved by increasing N_u and using the reference model to remove the overshoot. This is illustrated in Fig. 10 where $N_u = 2$ and $R_d = 1.5s + 1$. As seen from the figure the response is much improved.

Note that the choice of $N_u = N_y + p$ is not possible (when $\lambda = 0$) for the non-minimum phase systems since it removes the system zeros. So for a non-minimum phase system N_u must be $N_y + N_y + p$ (when $\lambda = 0$).

5.1.4. Time delay systems. Example 3 (a double integrator with unit time delay) was simulated to illustrate the ability of the CGPC to control time delay systems. A second order Pade approximation of the time delay was incorporated into A and B polynomials. CGPC design

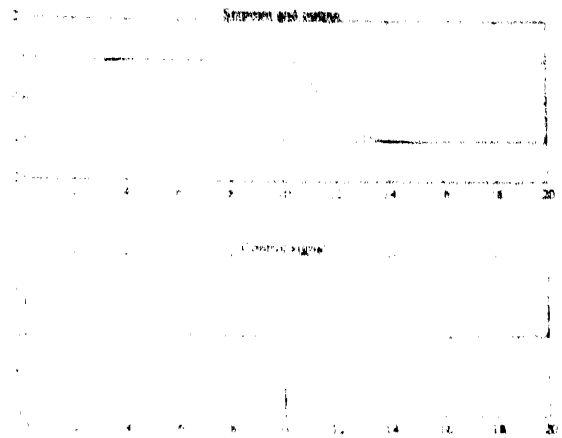


FIG. 10. Example 2 with $N_u = 2$, $R_d = 1.5s + 1$

was based on the resulting 4th order approximate system without time delay, but simulation was performed with the exact time delay. The CGPC parameters used in the simulation are as follows

$$\begin{aligned} R_u &= 1 \\ R_d &= 0.6s + 1 \\ T_1 &= 1 \\ T_2 &= 2 \\ N_y &= 20 \\ N_u &= 0 \\ \lambda &= 0. \end{aligned}$$

In the simulation, the sample interval was chosen to be 0.05s; the result is shown in Fig. 11. As can be seen from the figure, control performance is good, but this performance is obtained at the expense of increased N_y . However, this increase in N_y did not increase the computational burden significantly since $N_u = 0$. Although T_2 is chosen equal to the system time delay, it is also possible to obtain good control

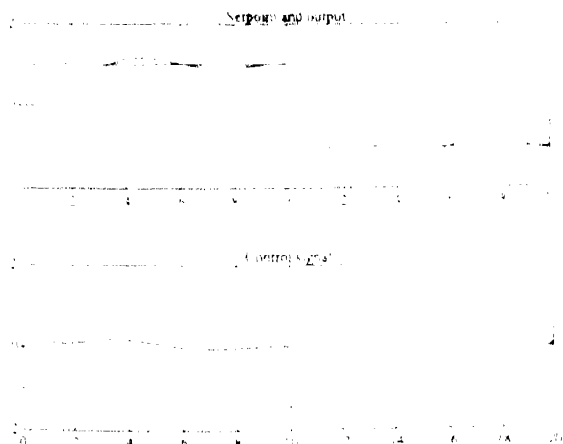


FIG. 9. Control of a non-minimum phase system

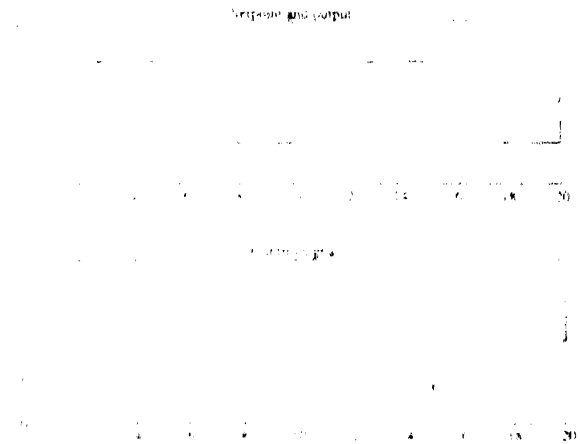


FIG. 11. Control of a time delay system

when $T_1 = 0$. However, in general, choosing T_1 equal to the system delay improves the control performance.

5.1.5. *LQ control.* As discussed in Section 3.4, the following setting of the CGPC parameters gives LQ control.

$$\begin{aligned} R_n/R_d &= 1 \\ N_u &= N_v - \rho \\ N_v &\rightarrow \infty \\ T_2 &\rightarrow \infty \\ T_1 &= 0. \end{aligned}$$

A double integrator [Example 3 without time delay and $C(s) = s + 1$] with the following choice of CGPC parameters was employed to illustrate this relationship.

$$\begin{aligned} N_v &= 12 \\ N_u &= 10 \\ \lambda &= 1. \end{aligned}$$

The closed-loop LQ poles for the above choice of λ are found as: $-0.7071 \pm 0.7071i$.

The closed-loop CGPC poles were calculated for $T_2 = 1$ to $T_2 = 10$ to see whether they converge to the LQ poles. The resulting poles are given below.

$$\begin{aligned} T_2 = 1 & \quad -0.1546 \pm 0.6654i \\ & \quad -0.6021 \pm 0.7920i \\ & \quad -0.6941 \pm 0.6987i \\ & \quad -0.6936 \pm 0.7002i \\ & \quad -0.7030 \pm 0.7071i \\ & \quad -0.7065 \pm 0.7075i \\ & \quad -0.7071 \pm 0.7071i \\ & \quad -0.7071 \pm 0.7071i \\ & \quad -0.7071 \pm 0.7071i \\ T_2 = 10 & \quad -0.7070 \pm 0.7072i \end{aligned}$$

As seen from the CGPC poles for $T_2 > 6$ CGPC poles and LQ poles become the same.

5.2. Parameter estimation

Any control approach can be combined with a recursive estimator to give its self-tuning version. In the self-tuning version of the CGPC, parameters of A and B polynomials are estimated by using the continuous-time least squares with forgetting.[†] Estimation of continuous-time system parameters can be performed by using a discrete-time estimator,

however for a continuous-time system a continuous-time estimation algorithm naturally seems more appropriate. Moreover, Gawthrop (1987) points out that discrete-time least squares is an approximation to the continuous-time one and so there are probably better ways of approximating the continuous-time least squares and thus obtaining better estimates.

In discrete-time least squares, the *inverse* of the information matrix, so-called covariance matrix, is updated. For numerical reasons, updating is performed by factoring the covariance matrix and updating the factors; such as the square-root or *U-D* factorization algorithms (Bierman, 1977). These methods guarantee that the covariance matrix always remains positive definite and thus non-singular. However, in CGPC there is no requirement on the system order and there may be situations where the system is overspecified. In this case the estimates are not unique (any common factors together with the actual parameters will be a solution to the estimation problem) and thus the information matrix is singular. Despite this, the above methods will try to update the inverse of a singular matrix which does not exist. In contrast, in our simulations, we update the information matrix itself and then take its pseudoinverse (Lawson and Hanson, 1974). This gives a unique solution which has a minimum Euclidean length among the other solutions. Of course this is not a numerically efficient algorithm but it is numerically stable.

Further work is needed to elucidate the fundamental and numerical problems arising from the essentially singular information matrix arising from the overspecified problem.

5.3. Adaptive simulations

In the first part of the simulations (Section 5.1), properties of the non-adaptive CGPC and effects of the CGPC design parameters were illustrated. Since these properties will remain the same in the self-tuning case, in this section we did not reconsider them; instead we simulated several examples in order to illustrate properties of the self-tuning CGPC. In the simulations, parameter \hat{a}_0 of the highest order s term of \hat{A} was fixed to 1 and thus one less A parameters were estimated. The following are common in the simulations.

1. All simulations start with a set of wrong parameters.
2. Estimator parameters: forgetting factor and initial inverse covariance are 0.2 and $0.00001I$ (where I is the unit matrix) respectively.

[†] We will not give the details of the algorithm here; more details appear in (Gawthrop 1987).

3. Sample interval is 0.05 sec.
4. Each figure consists of four graphs: the first one is the setpoint and output (some include model output as well), the second the control signal, and the third and fourth the estimated A and B parameters.

5.3.1. An example. A non-minimum phase system (Example 2) was simulated to give an example for the self-tuning CGPC. Simulation was performed with the same CGPC parameters as in the non-adaptive simulation corresponding to Figure 10. Two B and three A parameters, the same number as actual parameters, were estimated. Simulation result is shown in Fig. 12. As can be seen from the figure, parameters rapidly converge to their true values. Much more rapid convergence can be obtained if the initial inverse covariance is chosen zero. However, this results in large variations in the parameters at the beginning and this may give rise to initially large output. Note that, after convergence, the output is the same for adaptive and non-adaptive simulations as expected.

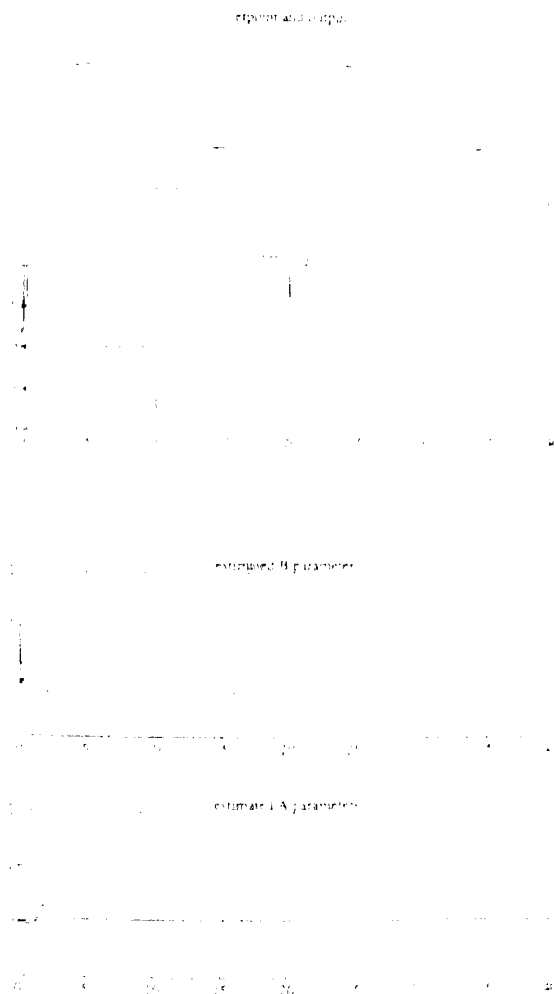


FIG. 12. Self-tuning CGPC (Example 2)

5.3.2. Effect of the noise. In practice, control systems are subject to disturbances of all kinds, such as stepwise load disturbances, high frequency sensor or thermal noise. In self-tuning control, although the underlying design method may have a good setpoint response and disturbance rejection, these disturbances may give rise to wrong parameter estimates and thus result in a bad control performance or even an unstable system. In general this problem can be avoided if the signal-to-noise ratio is high.

In order to see the performance of the CGPC when noises are present, we simulated Example 2 with an added random disturbance (Gaussian white noise with zero-mean and a standard deviation of 0.1) direct at the output. In the simulation exactly the same parameters as in Section 5.3.1 were used. The simulation result is shown in Fig. 13. As can be seen from the figure despite the noise parameter estimates and the control performance is good.

There are some fluctuations in the estimated parameters after time 40 sec. This appears to be due to the exponential forgetting factor causing the earlier part of the data (where the signal is

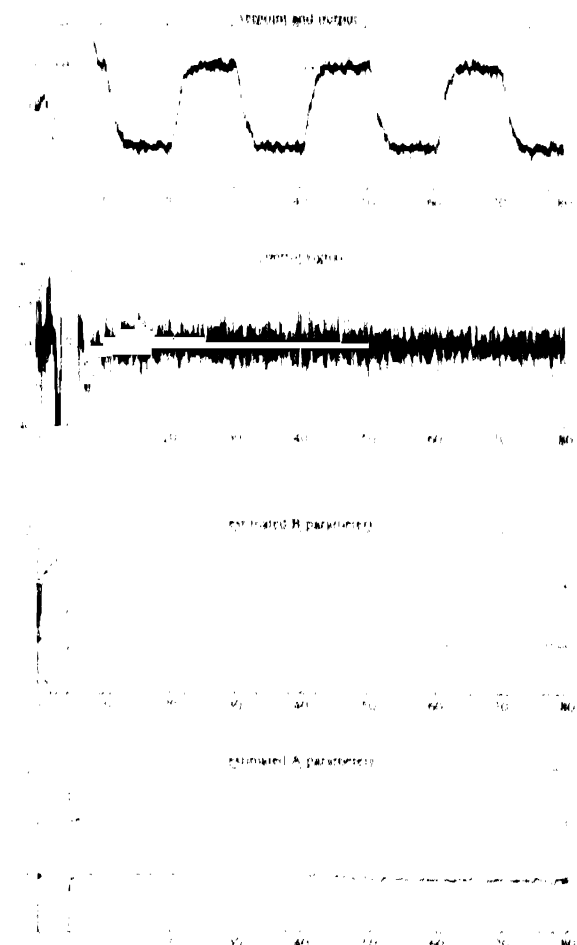


FIG. 13. Example 2 with added random disturbances

large compared to the noise) to be forgotten. The fluctuations are then due to the lower signal to noise ratio in the latter part of the simulation.

5.3.3. Time delay systems. In non-adaptive CGPC, time delay systems are approximated by a higher-order system without a delay. This is done by approximating the delay and incorporating it into system polynomials. The CGPC design is then based on the resulting approximate system. However, in the adaptive case, knowledge of the time delay is not available. One possible approach to this problem is to estimate the delay together with the system parameters (Besharati-Rad, 1988). Then the above design procedure can be applied. However, there are two drawbacks of this method: first, it will increase the complexity of the estimation algorithm, and second, each time instant we need to approximate the delay and then incorporate it into system polynomials to obtain the approximate system. The second approach, which removes these two drawbacks, is to let the estimator obtain the approximate system by specifying a higher-order model to the estimator. This is the approach taken in the self-tuning CGPC. Note that this approach

resembles the approach in discrete GPC where the time delay is taken into account by estimating a higher order B polynomial. Here we take the delay into account by estimating higher-order B and A polynomials.

Example 3 was simulated to illustrate the control of time delay systems by self-tuning CGPC. The CGPC design parameters and polynomials were chosen as in non-adaptive simulation (Section 5.1.4) except that R_d was chosen here $R_d = 0.8s + 1$ since $R_d = 0.6s + 1$ gave very slight overshoot. Four A and four B parameters were estimated. Simulation result is shown in Fig. 14. As can be seen from the figure control, performance is very good. Note that estimated system is non-minimum phase as expected.

5.3.4. Over parameterization. There may be some situations where the parameters of the model are overspecified, this results in common factors in the estimated model. Pole-placement algorithms (Wellstead *et al.*, 1979) fail under these circumstances. In Section 3.3 the effect of common factors is examined and shown that, as for the GPC it is not a problem for the CGPC. This will also be illustrated by simulation.

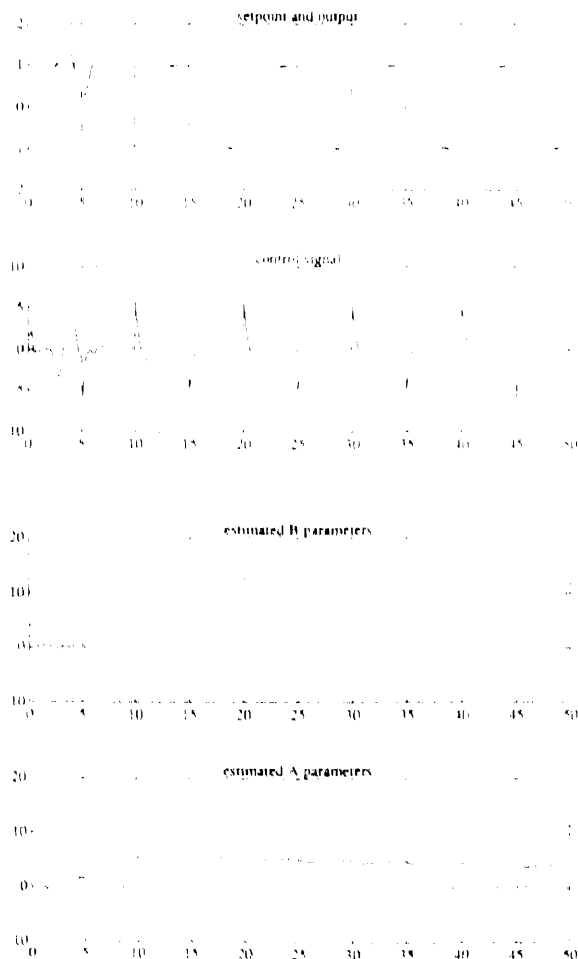


FIG. 14. Self-tuning control of a time delay system.

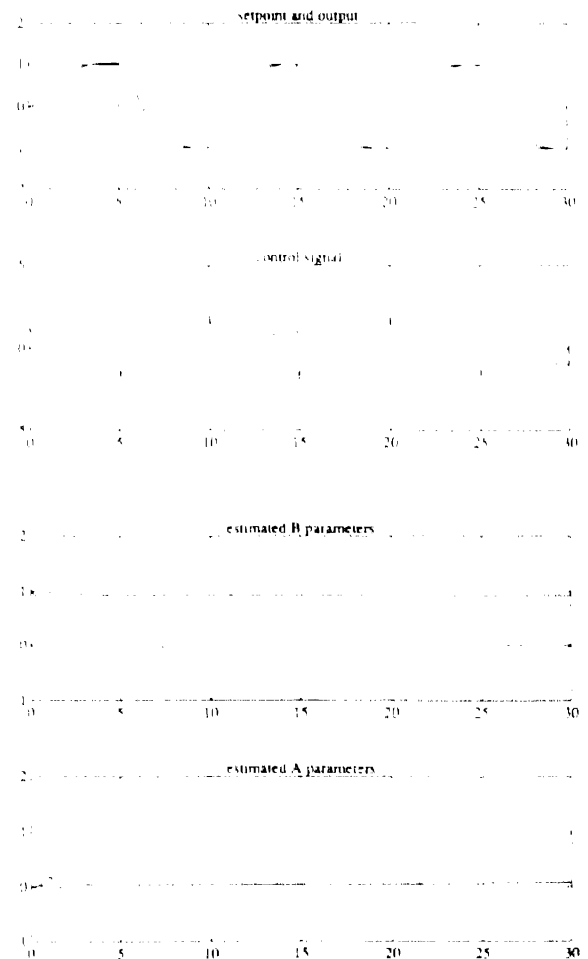


FIG. 15. The effect of over parameterization.

Example 4 was simulated for this purpose with the following CGPC parameters:

$$R_n = 1$$

$$R_d = 1.5s + 1$$

$$T_1 = 0$$

$$T_2 = 1$$

$$N_u = 6$$

$$N_u = 0$$

$$\lambda = 0.$$

One more each of A and B (3A and 2B) parameters were estimated. The simulation result is shown in Fig. 15. As can be seen from the figure there is a common factor at $s = 0$ in the estimated system model. This common factor did not affect the control performance. If the simulation is repeated with exact parameterization, it can be seen that the same output response is obtained.

6. CONCLUSIONS

In this paper the CGPC, a continuous-time analogue of the GPC developed by Clarke *et al.*, (1987), is presented. Like its discrete-time counterpart, the CGPC algorithm appears to be a useful design method. A large variety of systems can be controlled reasonably well by only adjusting the prediction horizons T_1 and T_2 keeping the control order N_u as zero but, for the more complex systems (at the same time higher-order, open-loop unstable, non-minimum phase) an increased value of N_u is needed. In general, it is advisable to keep N_u small in order not to increase the computational burden, instead use reference-model R_n/R_d to adjust the transients. Apart from this R_n/R_d can also be used to obtain model-following type control (sometimes exact model-following) with a large N_u .

The relationship of CGPC with LQ control is also examined, and it is shown that LQ control can be considered as a subalgorithm of the CGPC. In addition, it is shown that time-delay systems can be controlled in a similar way to GPC by increasing the system order in order to accommodate a rational approximation to the delay.

An important feature of the CGPC algorithm is that, unlike polynomial LQ and pole-placement controllers, it is not necessary to solve a Diophantine equation to compute the control law. In particular, the case of systems with cancelling pole/zero pairs is considered and shown to cause no difficulty. In addition, unlike the GMV algorithm, control weighting is not necessary for the control of non-minimum phase systems.

The CGPC algorithm can also be further generalized by using an auxiliary output approach and dynamic control weighting. This will probably make it possible to consider the model-reference and pole-placement control (and also their detuned versions) in the CGPC frame work.

In these respects, the CGPC is superficially similar to its discrete-time counterpart, but there are important differences in the way in which output prediction and control weighting is accomplished. An example of this is that, whereas the GPC constrains the predicted control difference to be zero after N_u samples, the CGPC constrains the predicted control so that the derivatives of order greater than N_u are zero.

In brief, CGPC (as GPC) appears to be a very suitable control algorithm for the self-tuning control applications for a large variety of systems.

Acknowledgement. We are grateful to the Turkish Ministry of Education for providing financial support to the first author during the course of study reported here.

REFERENCES

- Astrom, K. J. (1970) *Introduction to Stochastic Control Theory*. Academic Press, New York.
- Besharati Rad, A. (1988) Identification and adaptive control of retarded systems: A continuous time approach. PhD Thesis, School of Engineering and Applied Sciences, University of Sussex, U.K.
- Bierman, G. J. (1977) *Factorization Methods for Discrete Sequential Estimation*. Academic Press, New York.
- Clarke, D. W. and P. J. Gawthrop (1975) Self tuning controller. *IEE Proc. Pt. D* **122**, 929-934.
- Clarke, D. W. and P. J. Gawthrop (1979) Self tuning control. *IEE Proc. Pt. D* **126**, 635-640.
- Clarke, D. W., C. Mohtadi and P. S. Tuffs (1987) Generalised Predictive Control: Part 1: the basic algorithm and Part 2: extensions and interpretations. *Automatica* **23**, 137-160.
- Clarke, D. W. and L. Zhang (1987) Long range predictive control using weighting sequence models. *IEE Proc. Pt. D* **134**, 187-195.
- Cutler, C. R. and B. J. Ramaker (1980) Dynamic matrix control: a computer control algorithm. *Proc. Joint Automatic Control Conf.*, San Francisco, U.S.A.
- Gawthrop, P. J. (1986) Continuous time self-tuning control: A unified approach. *Preprints 2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing*, Lund, Sweden, pp. 19-24.
- Gawthrop, P. J. (1987) *Continuous time Self tuning Control*. Research Studies Press, Letchworth, U.K.
- Kalath, T. (1980) *Linear Systems*. Prentice Hall, Englewood Cliffs, N.J.
- de Keyser, R. M. C. and A. R. van Cauwenberghe (1981) A self-tuning multistep predictor application. *Automatica* **17**, 167-174.
- de Keyser, R. M. C. and A. R. van Cauwenberghe (1985) Extended prediction self-adaptive control. *Proc. 7th IFAC/IFORS Symp. on Identification and System Parameter Estimation*, York, U.K.
- de Keyser, R. M. C., Ph. G. A. van de Velde and E. A. G. Dumortier (1988) A comparative study of self-adaptive long range predictive control methods. *Automatica* **24**, 149-163.
- Lawson, C. L. and R. J. Hanson (1974) *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, N.J.

- Lehi, M. A. and M. B. Zarrop (1987). Generalized pole-placement self-tuning controller, Part 1. Basic algorithm. *Int. J. Control* **46**, 547–568.
- Marshall, J. E. (1979). *Control of time-delay systems*. Peter Peregrinus, Stevenage, U.K.
- Mosca, E., G. Zappa and C. Manfredi (1984). Multistep horizon self-tuning controllers: the MUSMAR approach. *Preprints IFAC 9th World Congress*, Budapest, Hungary, Vol. VII, pp. 155–160.
- Peterka, V. (1984). Predictor-based self-tuning control. *Automatica* **20**, 39–50.
- Richalet, J., A. Rault, J. L. Testud and J. Papon (1978). Model predictive heuristic control: applications to industrial processes. *Automatica* **14**, 413–428.
- de Souza, C. E., G. C. Goodwin, D. O. Mayne and M. Palaniswami (1988). An adaptive control algorithm for linear systems having unknown time delay. *Automatica* **24**, 327–341.
- Wellstead, P. E., D. Prager and P. Zanker (1979). Pole assignment self-tuning regulator. *IEE Proc. Pt. D* **126**, 781–787.
- Wolovich, W. A. (1974). *Linear multivariable systems*. Springer, Berlin.
- Ydstie, B. E. (1984). Extended horizon adaptive control. *Preprints IFAC 9th World Congress*, Budapest, Hungary, Vol. VII, pp. 133–138.

Robust Gamma-stability Analysis in a Plant Parameter Space*

J. ACKERMANN,^{††} D. KAESBAUER[†] and R. MUENCH[†]

Polynomial parameter dependency of the characteristic polynomial of a control system can be handled by constructing stability and performance regions in the parameter space. This robustness analysis is applied to automatic steering of a bus.

Key Words—Robust control, stability, control system analysis, nonlinear equations, control applications, automobiles

Abstract—Given a characteristic polynomial whose coefficients depend polynomially on l uncertain parameters, the following robustness problem arises: Determine whether all the roots of the polynomial are located in a prescribed region Γ in the complex plane for all admissible parameter values. To this end, the boundary $\partial\Gamma$ of Γ is mapped into the parameter space. A necessary and sufficient condition for Γ -stability of an operating domain in parameter space is that it contains at least one Γ -stable point and is not intersected by the image of $\partial\Gamma$. This condition may be tested graphically by gridding $l+2$ parameters and projecting all boundaries into a two-dimensional subspace of the parameter space. Finally the method is applied to a track guided bus with uncertain mass and velocity.

1. INTRODUCTION

CONSIDER A linear, time-invariant system with characteristic polynomial

$$P(s, \mathbf{q}) = a_0(\mathbf{q}) + a_1(\mathbf{q})s + \dots + a_n(\mathbf{q})s^n \quad (1)$$

where $\mathbf{q} = [q_1, q_2, \dots, q_l]'$ is a vector of real uncertain parameters in a set Q of allowable perturbations. $\mathbf{a}(\mathbf{q}) = [a_0(\mathbf{q}), a_1(\mathbf{q}), \dots, a_n(\mathbf{q})]'$ is the real coefficient vector of the polynomial (1). Assume $a_n(\mathbf{q}) > 0$ for all $\mathbf{q} \in Q$ and let $s = \sigma + j\omega$.

Given Γ , a subset of the complex plane, and Q

we are interested in a yes/no answer to the following robust stability question: Is $P(s, \mathbf{q})$ Γ -stable for all $\mathbf{q} \in Q$? That is, does $P(s, \mathbf{q})$ have all its roots in Γ for all $\mathbf{q} \in Q$?

Noting that analysis tools are frequently used for design by trial and error, it is desirable to gain insight from the analysis beyond a yes/no answer. In this regard, one wants to know how small Γ can be made, how large Q can be made or how controller parameters must be changed such that robust stability for a given Γ and Q is achieved. These questions motivate the approach of this paper which is aimed at a study of the conflicts and possible tradeoffs between Γ and Q .

The difficulty of the robust stability analysis above depends on the kind of coefficient functions $\mathbf{a}(\mathbf{q})$, the number l of parameters, the shapes of Q and Γ and on the polynomial degree n .

Frazer and Duncan (1929) have shown that for continuous coefficient functions $\mathbf{a}(\mathbf{q})$ a necessary and sufficient condition for robust stability is that (i) there exists a $\mathbf{q} = \mathbf{q}_0 \in Q$ such that $P(s, \mathbf{q}_0)$ is stable; and (ii) $P(s, \mathbf{q})$ does not have roots on the imaginary axis for any $\mathbf{q} \in Q$.

Condition (i) is easily tested by checking the stability of $P(s, \mathbf{q}_0)$ for an arbitrary $\mathbf{q}_0 \in Q$. If $P(s, \mathbf{q}_0)$ is unstable, then we cannot have robust stability, if $P(s, \mathbf{q}_0)$ is stable, then condition (i) is satisfied. Frazer and Duncan (1929) have also shown that condition (ii) is satisfied if and only if $P(s, \mathbf{q})$ neither has a real root at $s = 0$, i.e.

$$a_0(\mathbf{q}) \neq 0 \quad \text{for all } \mathbf{q} \in Q \quad (2)$$

nor an imaginary pair of roots at $s_{1,2} = \pm j\omega$, i.e.

$$\det \mathbf{H}_{n-1}(\mathbf{q}) \neq 0 \quad \text{for all } \mathbf{q} \in Q \quad (3)$$

where \mathbf{H}_{n-1} is the last but one Hurwitz matrix,

* Received 13 March 1989; revised 13 November 1989; received in final form 22 February 1990. The original versions of this paper were presented at the 8th IFAC World Congress on Control Science and Technology for the Progress of Society which was held in Kyoto, Japan during August 1981 and the 10th IFAC World Congress which was held during July 1987. The Published Proceedings of these IFAC Meetings may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford, OX3 0BW U.K. This paper was recommended for publication in revised form by Associate Editor H. Kimura under the direction of Editor H. Kwakernaak.

[†]Laboratory for Flight Systems Dynamics, German Aerospace Research Organisation (DLR), D-8031 Oberpfaffenhofen, F.R.G.

^{††}Author to whom all correspondence should be addressed.

i.e.

$$\mathbf{H}_{n-1}(\mathbf{q}) = \begin{bmatrix} a_1(\mathbf{q}) & a_2(\mathbf{q}) & \cdot & \cdot & 0 \\ a_0(\mathbf{q}) & a_2(\mathbf{q}) & \cdot & \cdot & \cdot \\ 0 & a_1(\mathbf{q}) & a_3(\mathbf{q}) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & a_{n-1}(\mathbf{q}) \end{bmatrix}. \quad (4)$$

Note that the case of a root crossing the stability boundary at $s = \infty$ has been excluded by the assumption $a_n(\mathbf{q}) > 0$. In fact we could have divided $P(s, \mathbf{q})$ by $a_n(\mathbf{q})$ in order to make the polynomial monic. It is easier however to deal with polynomial coefficient functions $a_0(\mathbf{q}) \cdots a_n(\mathbf{q})$ than with rational coefficient functions $a_0(\mathbf{q})/a_n(\mathbf{q}) \cdots a_{n-1}(\mathbf{q})/a_n(\mathbf{q})$. Thus a robust stability analysis for the case of polynomial coefficient functions covers the case of monic polynomials with rational coefficient functions.

In the recent literature on robustness analysis studies were made how to make the "for all $\mathbf{q} \in Q$ " condition tractable. Typically Q is assumed in form of a box

$$q_i \in [q_i^-; q_i^+], \quad i = 1, 2, \dots, l. \quad (5)$$

Kharitonov (1978) proved, that for interval polynomials, i.e. $a_i = q_i$, it suffices to check four specific vertices of the box (5) for stability. For affine coefficient functions $a_i(\mathbf{q}) = a_{i0} + \mathbf{c}_i^T \mathbf{q}$, Bartlett *et al.* (1988) have shown that it suffices to check the exposed edges of Q for stability. An edge $\mathbf{q}_A - \mathbf{q}_B$ can be tested by Hurwitz matrices of $P(s, \mathbf{q}_A)$ and $P(s, \mathbf{q}_B)$, [see Bialas (1985)]. A simplified version of this test was given by Ackermann and Barmish (1988): The edge $\mathbf{q}_A - \mathbf{q}_B$ is stable if and only if $\mathbf{H}_{n-1}(\mathbf{q}_A)\mathbf{H}_{n-1}(\mathbf{q}_B)$ does not have negative real eigenvalues. For multilinear coefficient functions, sufficient conditions for robust stability are available (Zadeh and Desoer, 1963) and necessary and sufficient conditions are obtained only under restrictive assumptions.

For the analysis of systems with general multilinear or polynomial coefficient functions in the characteristic polynomial, one of the possibilities is brute force gridding of Q . It is possible however to use the above results in order to reduce the gridding dimensionality and computational effort per grid point (Ackermann *et al.*, 1988).

The Frazer–Duncan condition (ii) can be tested in different ways:

1. Zero exclusion from the value set. Here the condition (ii) is interpreted as $P(j\omega, \mathbf{q}) \neq 0$ for all ω and all $\mathbf{q} \in Q$. The frequency ω is gridded and for each frozen frequency $\omega = \omega^*$ the complex set

$$P_{\omega^*} = \{P(j\omega^*, \mathbf{q}) : \mathbf{q} \in Q\} \quad (6)$$

is constructed. (This construction may require gridding of Q .) P_{ω^*} must not contain the point zero [see the survey by Barmish (1988) and the references therein].

2. Parameter space approach. Here the set of all real \mathbf{q} is investigated such that $P(s, \mathbf{q})$ has roots on the imaginary axis.

$$Q_{j\omega} = \{\mathbf{q} : P(j\omega, \mathbf{q}) = 0 \text{ for some } \omega \geq 0\}. \quad (7)$$

A necessary and sufficient condition for stability robustness is that (i) there exists a $\mathbf{q}_0 \in Q$ such that $P(s, \mathbf{q}_0)$ is stable and (ii) $Q_{j\omega}$ does not intersect Q .

$P(j\omega, \mathbf{q})$ may be written as

$$P(j\omega, \mathbf{q}) = U(\omega^2, \mathbf{q}) + j\omega V(\omega^2, \mathbf{q}) \quad (8)$$

where

$$U(\omega^2, \mathbf{q}) = a_0(\mathbf{q}) - a_2(\mathbf{q})\omega^2 + a_4(\mathbf{q})\omega^4 - \dots \quad (9)$$

$$V(\omega^2, \mathbf{q}) = a_1(\mathbf{q}) - a_3(\mathbf{q})\omega^2 + a_5(\mathbf{q})\omega^4 - \dots \quad (10)$$

For $\omega = 0$ the real root boundary $Q_{Rc} = \{\mathbf{q} : a_0(\mathbf{q}) = 0\}$ is obtained. For $\omega > 0$ the set $Q_{Im} = \{\mathbf{q} : U(\omega^2, \mathbf{q}) = 0 \text{ and } V(\omega^2, \mathbf{q}) = 0 \text{ for some } \omega > 0\}$ must be described, then $Q_{j\omega} = Q_{Rc} \cup Q_{Im}$.

Frazer and Duncan (1929) apply the Sylvester test to the polynomials U and V for relative primeness. This results in a condition

$$\det \mathbf{H}_{n-1}(\mathbf{q}) = 0 \quad (11)$$

for U and V to have a common root, where $\mathbf{H}_{n-1}(\mathbf{q})$ is given in equation (4). Thus ω is eliminated. By Orlando's (1911) formula, (Gantmacher, 1959)

$$\det \mathbf{H}_{n-1}(\mathbf{q}) = (-1)^{n(n-1)/2} (a_n(\mathbf{q}))^{n-1} \prod_{i=1}^{n-1} \prod_{k=i+1}^n (s_i(\mathbf{q}) + s_k(\mathbf{q})) \quad (12)$$

where $s_i(\mathbf{q})$, $s_k(\mathbf{q})$ are roots of $P(s, \mathbf{q})$. Obviously $\det \mathbf{H}_{n-1}(\mathbf{q}) = 0$ for $s_i(\mathbf{q}) = j\omega$, $s_k(\mathbf{q}) = -j\omega$; i.e. Q_{Im} is a subset of $Q_H = \{\mathbf{q} : \det \mathbf{H}_{n-1}(\mathbf{q}) = 0\}$. But $\det \mathbf{H}_{n-1}$ also vanishes for a pair of real roots $s_i(\mathbf{q}) = a$, $s_j(\mathbf{q}) = -a$. (In the elimination of ω the restriction to real values of ω got lost.) The values of \mathbf{q} that are contained in Q_H , but not in Q_{Im} , give rise to unstable polynomials. Thus for robust stability also Q_H must not intersect Q .

In this paper we derive a description of Q_{im} by solving the equations $U(\omega^2, \mathbf{q}) = 0$, $V(\omega^2, \mathbf{q}) = 0$ for

$$q_1 = q_1(q_3, \dots, q_l, \omega) \quad (13)$$

$$q_2 = q_2(q_3, \dots, q_l, \omega). \quad (14)$$

ω is not eliminated, thus the restriction to real values $\omega > 0$ is preserved. A strength of the approach considered here is its ability to handle general polynomial coefficient functions, i.e. $\mathbf{a}(\mathbf{q})$ and therefore $U(\omega^2, \mathbf{q})$ and $V(\omega^2, \mathbf{q})$ may be polynomial in q_1 and q_2 and arbitrary continuous in q_3, q_4, \dots, q_l . q_1 and q_2 are calculated for each admissible grid point $q_3^* \dots q_l^*, \omega^*$. Real values q_1 and q_2 obtained from (13) and (14) must be located outside the operating domain, i.e.

$$q_1 \notin [q_1^-; q_1^+] \quad (15)$$

or

$$q_2 \notin [q_2^-; q_2^+].$$

Thus two parameters q_1 and q_2 can be treated as continuous variables, only further parameters must be gridded.

In the next section the concept is applied to a simple example with affine coefficient functions. In Section 3 the approach is generalized to other regions Γ in the complex plane. Polynomial coefficient functions are treated in Section 4. The example of a track-guided bus shows the practical application of the method.

2. AN EXAMPLE WITH AFFINE COEFFICIENT FUNCTIONS

Consider

$$\begin{aligned} P(s, q_1, q_2, q_3) = & (12 - 4q_1 - q_2 + q_3) \\ & + (-44 + 19q_1 + 8q_2 + q_3)s \\ & + (78 - 9.625q_1 - 16q_2)s^2 \\ & + (-24 + 16q_1)s^3 + 16s^4 \end{aligned} \quad (16)$$

$$q_1 \in [2; 2.5], \quad q_2 \in [1; 2], \quad q_3 \in [0; 3]$$

$$U(\omega^2, q_1, q_2, q_3) = (12 - 4q_1 - q_2 + q_3) - (78 - 9.625q_1 - 16q_2)\omega^2 + 16\omega^4 = 0$$

$$\begin{aligned} V(\omega^2, q_1, q_2, q_3) = & (-44 + 19q_1 + 8q_2 + q_3) \\ & - (-24 + 16q_1)\omega^2 = 0. \end{aligned} \quad (17)$$

Solve these two linear equations for q_1 and q_2 :

$$q_1 = \frac{4(64\omega^4 + 4q_3\omega^2 - 26\omega^2 - 2.25q_3 - 13)}{256\omega^4 - 243\omega^2 - 13}$$

q_2 :

$$\frac{512\omega^6 - 2642\omega^4 + 51.25q_3\omega^2 + 2309\omega^2 - 46q_3 - 104}{2 \cdot (256\omega^4 - 243\omega^2 - 13)}$$

Singularities of these curves can occur at frequencies where the denominator vanishes i.e. $\omega_1 = 1$. For ω_1 the stability boundary in the (q_1, q_2) -plane has an asymptote, because (17) represents two parallel but not identical lines. If q_3 is chosen such that the numerator of the expressions for q_1 or q_2 also vanishes at ω_1 , then an entire straight line belongs to the stability boundary for $\omega = \omega_1$. In the example this occurs for $q_3^{(1)} = 0.8929$ and the equation of the straight line is given by $U(1, q_1, q_2, q_3^{(1)}) = \text{constant} \neq V(1, q_1, q_2, q_3^{(1)}) = 0$. This value is outside the operating domain.

For each fixed $q_3 = q_3^* \in [0, 3]$ and each fixed $\omega = \omega^*$ a point $q_1(q_3^*, \omega^*)$, $q_2(q_3^*, \omega^*)$ is obtained. For stability robustness all these points must satisfy $[q_1, q_2] \notin [[2; 2.5], [1; 2]]$.

For a graphical representation in the (q_1, q_2) -plane a reasonable stepsize in ω must be chosen (like in Nyquist plots) and for each q_3^* a continuous stability boundary is obtained. All these boundaries are projected into the same (q_1, q_2) -plane. The projection of the three-dimensional Q -box is a rectangle. Figure 1 shows the complex root boundaries (CRB) and the real root boundaries (RRB). $a_0(\mathbf{q}) = 12 - 4q_1 - q_2 + q_3 = 0$ for $q_3^* = 0, 0.75, 1.5, 2.25$ and 3.

It is seen that the original operating domain is not entirely stable. For $q_3 = 3$ the complex root boundary intersects Q for frequencies $\omega \in [1.16; 1.24]$. We could for example try a compensation for this frequency band or reduce Q to $q_1 \in [2; 2.5]$, $q_2 \in [1; 1.8]$ in order to obtain a stable operating domain.

For other examples with additional parameters q_4, \dots, q_l an $l - 2$ dimensional grid for q_3, q_4, \dots, q_l must be chosen. For each grid point the stability boundary may be projected into the (q_1, q_2) -plane and compared there with the rectangle $q_1 \in [q_1^-, q_1^+]$, $q_2 \in [q_2^-, q_2^+]$.

3. GENERALIZATION TO l STABILITY REGIONS

The designer of a control system is not satisfied with mere stability for the given operating domain; he wants to achieve performance. One aspect of performance can easily be captured in the present context, that is the location of the closed-loop eigenvalues in a subset of the complex plane. Thus properties like minimum damping or negative real part can be assured. Also the designer may have done a design for nominal parameter values and wants

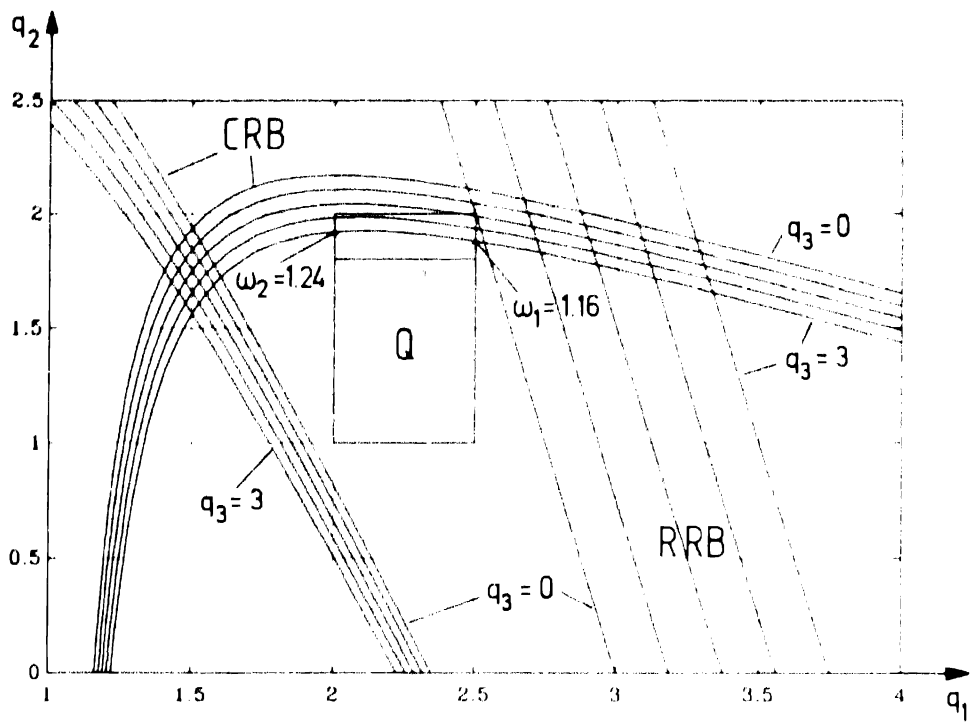


FIG. 1. Complex (CRB) and real root boundaries (RRB) for the example (16). Boundaries for $q_3 = 0, 0.75, 1.5, 2.25$ and 3 are projected into the (q_1, q_2) -plane. The projection of the Q -Box is a rectangle. The system is stable for $q_1 \in [2, 2.5], q_2 \in [1; 1.8], q_3 \in [0, 3]$.

to guarantee that poles do not move too far from their nominal positions under the influence of parameter variations. Figure 2 shows an example, where three eigenvalues have separate allowed variations and all other eigenvalues should remain to the left of a further boundary.

The continuity argument of Frazer and Duncan (1929) is valid also for this case: If \mathbf{q} varies continuously in Q , then also $\mathbf{a}(\mathbf{q})$ and the roots of $P(s, \mathbf{q}) = [1, s \cdots s^n] \mathbf{a}(\mathbf{q})$ vary continuously. The roots are not allowed to cross the

boundary $\partial\Gamma$ of the Γ -stability region as they move away from a Γ -stable starting position. This is equivalent to saying that the image $Q_{\partial\Gamma}$ of $\partial\Gamma$ in \mathbf{q} -space must not intersect Q .

Like in the case of the left half plane $Q_{\mathcal{A}}$ is composed of real root and complex root boundaries: $Q_{\mathcal{A}} = Q_{\text{Im } 1} \cup Q_{\text{Im } 2} \cup Q_{\text{Re } 1} \cup Q_{\text{Re } 2} \cdots$. In the example of Fig. 2 the real root boundaries are given by

$$Q_{\text{Re } i} = \{\mathbf{q}; P(\sigma_i, \mathbf{q}) = 0\}, \quad i = 1, 2, 3. \tag{18}$$

The set Γ is a union of two complex and one real set, thus the complex boundary is a union of two branches. In other examples Γ may be the intersection of two sets, e.g. damping greater than $1/\sqrt{2}$ and $\text{Re } s_i < -1$. In this case the boundary $\partial\Gamma$ is defined and parameterized piecewise, in the example, with $\alpha = \omega$,

$$Q_{\text{Im } 1} = \{\mathbf{q}; P(-1 + j\alpha, \mathbf{q}) = 0 \text{ for some } \alpha \in [0; 1]\} \tag{19}$$

$$Q_{\text{Im } 2} = \{\mathbf{q}; P(-\alpha + j\alpha, \mathbf{q}) = 0 \text{ for some } \alpha \in [1; \infty)\}. \tag{20}$$

The image of the complex boundary is the union of the segments $Q_{\text{Im } 1} \cup Q_{\text{Im } 2}$.

In the following a computationally tractable

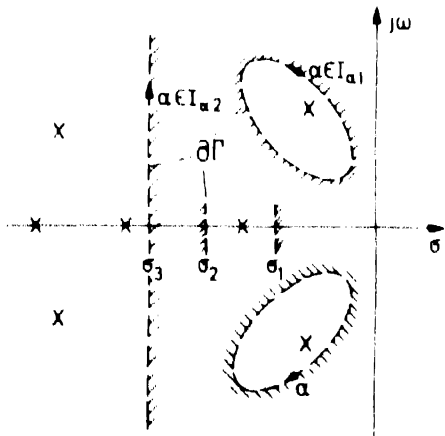


FIG. 2. Crosses indicate closed loop pole locations for nominal parameters $\mathbf{q} = \mathbf{q}_0$. Γ -stability is defined by boundaries around groups of eigenvalues.

description of

$$Q_{1m} = \{q: P(s, q) = 0 \text{ for some } s \in \partial\Gamma\} \quad (21)$$

is given.

In analogy to the frequency ω we define a scalar parameter α called "generalized frequency" that parameterizes the complex branches of $\partial\Gamma$. The boundary $\partial\Gamma$ may for example be described by a sufficiently dense list $\sigma(\alpha_i), \omega(\alpha_i), i = 1, 2, 3, \dots$ or by continuous functions $\sigma(\alpha), \omega(\alpha)$ on each branch or segment of $\partial\Gamma$ for α in an index set I_α , i.e. Q_{1m} may be written

$$Q_{1m} = \bigcup_{\alpha \in I_\alpha} Q_i(\alpha). \quad (22)$$

For a fixed $\alpha = \alpha^*$, i.e. a fixed complex conjugate pair $s(\alpha^*) = \sigma(\alpha^*) + j\omega(\alpha^*)$ on $\partial\Gamma$, $Q_{1m}(\alpha^*)$ is the set of parameter values q such that $P(s(\alpha^*), q) = 0$. $\partial\Gamma$ is symmetric with respect to the real axis.

Each branch of $\partial\Gamma$ must be either a finite closed contour or a contour extending to infinity, such that $\partial\Gamma$ provides a clear distinction between a Γ -stable and a Γ -unstable location for each root of $P(s, q)$.

The main result of this section is formulated as the following boundary representation theorem. A preliminary version of this theorem has appeared in a conference paper (Ackermann and Kaesbauer, 1981).

Boundary representation theorem

$q \in Q_{1m}(\alpha)$ if and only if

$$\begin{bmatrix} d_0(\alpha) & d_1(\alpha) & \cdots & d_n(\alpha) \\ 0 & d_0(\alpha) & \cdots & d_{n-1}(\alpha) \end{bmatrix} \mathbf{a}(q) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (23)$$

for some $\alpha \in I_\alpha$, where

$$\begin{aligned} d_0(\alpha) &= 1 \\ d_1(\alpha) &= 2\sigma(\alpha) \\ d_{i+1}(\alpha) &= 2\sigma(\alpha)d_i(\alpha) \\ &\quad - [\sigma^2(\alpha) + \omega^2(\alpha)]d_{i-1}(\alpha), \\ &\quad i = 1, 2, \dots, n-1 \end{aligned} \quad (24)$$

Proof. Consider the polynomial $P(s, q) = [1, s, \dots, s^n] \mathbf{a}(q)$ for a fixed $q = q^*$. It has a complex conjugate pair of roots on $\partial\Gamma$ at $\sigma(\alpha) \pm j\omega(\alpha)$, $\alpha \in I_\alpha$, if and only if

$$P(s, q^*) = [\sigma^2(\alpha) + \omega^2(\alpha) - 2\sigma(\alpha)s + s^2]R(s, q^*) \quad (25)$$

where

$$\begin{aligned} R(s, q^*) &= r_0 + r_1s + \cdots + r_{n-2}s^{n-2} \\ &= [1, s, \dots, s^{n-2}]r \end{aligned} \quad (26)$$

is an arbitrary polynomial with real coefficients. Equivalently (omitting the dependency of σ and ω on α for notational convenience)

$$\mathbf{a}(q)^* = \begin{bmatrix} 1 & -2\sigma & \sigma^2 + \omega^2 & 0 & \cdots & 0 \\ 0 & 1 & -2\sigma & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \sigma^2 + \omega^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & -2\sigma & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & \cdots & \cdots \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ r_0 \\ \vdots \\ r_{n-2} \end{bmatrix} \quad (27)$$

The matrix in (27) is triangular with identical elements on the diagonals. Therefore its inverse \mathbf{D} has the same structure. The entries d_i of \mathbf{D} are determined from

$$\begin{bmatrix} d_0 & d_1 & \cdots & d_n \\ 0 & d_0 & \cdots & d_{n-1} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & d_1 & \cdots \\ \cdots & \cdots & d_0 & \cdots \end{bmatrix} \begin{bmatrix} 1 & -2\sigma & \sigma^2 + \omega^2 & 0 & \cdots & 0 \\ 0 & 1 & -2\sigma & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \sigma^2 + \omega^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & -2\sigma & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & \cdots & \cdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix} \quad (28)$$

which forces

$$\begin{aligned} d_0 &= 1 \\ 2\sigma d_0 + d_1 &= 0 \\ (\sigma^2 + \omega^2)d_0 - 2\sigma d_1 + d_2 &= 0 \\ &\vdots \\ (\sigma^2 + \omega^2)d_{n-2} - 2\sigma d_{n-1} + d_n &= 0 \end{aligned} \quad (29)$$

$$\begin{aligned} d_0 &= 1 \\ d_1 &= -2\sigma \\ d_2 &= 2\sigma d_1 - (\sigma^2 + \omega^2)d_0 \\ &\vdots \\ d_n &= 2\sigma d_{n-1} - (\sigma^2 + \omega^2)d_{n-2} \end{aligned}$$

Premultiplying equation (27) by **D** we have $\mathbf{q}^* \in Q_{lm}(\alpha)$ if and only if

$$\begin{bmatrix} d_0 & d_1 & \cdots & d_n \\ 0 & d_0 & & \\ & & \ddots & \\ & & & d_1 \\ 0 & & 0 & d_0 \end{bmatrix} \mathbf{a}(\mathbf{q}^*) = \begin{bmatrix} 0 \\ 0 \\ r_0 \\ \vdots \\ r_{n-2} \end{bmatrix}. \quad (30)$$

The last $n-1$ rows of (30) have an undetermined right hand side because the remainder polynomial $R(s, \mathbf{q}^*)$ is arbitrary. There only remain the first two rows. Hence we conclude that $\mathbf{q}^* \in Q_{lm}(\alpha)$ if and only if

$$\begin{bmatrix} d_0(\alpha) & d_1(\alpha) & \cdots & d_n(\alpha) \\ 0 & d_0(\alpha) & \cdots & d_{n-1}(\alpha) \end{bmatrix} \mathbf{a}(\mathbf{q}^*) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (31)$$

The fixed value \mathbf{q}^* may be replaced by the general vector \mathbf{q} to obtain equation (23). Thus the proof of the theorem is complete.

By equation (22) we now conclude that $\mathbf{q} \in Q_{lm}$ if and only if there exists an $\alpha \in I_\alpha$ such that equation (23) holds. In other words, in analogy to the ω -sweep over the imaginary axis in the robust stability test of Section 2, an α -sweep along all branches or segments of $\partial\Gamma$ must be made. For each α the set $Q_{lm}(\alpha)$ can be calculated from (23).

Example 1. Imaginary axis, $\sigma(\alpha) = 0$, $\omega^2(\alpha) = \alpha$, $\alpha \in [0; \infty)$

$$\begin{aligned} d_0 &= 1 \\ d_1 &= 0 \\ d_{i+1} &= -\alpha d_{i-1} \end{aligned} \quad (32)$$

$$\begin{bmatrix} 1 & 0 & \cdots & \alpha & 0 & \alpha^2 & \cdots \\ 0 & 1 & 0 & -\alpha & 0 & \cdots \end{bmatrix} \mathbf{a}(\mathbf{q}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (33)$$

This expression is equivalent to (9) and (10).

Example 2. Unit circle, $\sigma(\alpha) = \alpha$, $\sigma^2(\alpha) + \omega^2(\alpha) = 1$, $\alpha \in [-1; 1]$

$$\begin{aligned} d_0 &= 1 \\ d_1 &= 2\alpha \\ d_{i+1} &= 2\alpha d_i - d_{i-1} \end{aligned} \quad (34)$$

$$\begin{bmatrix} d_0 & d_1 & d_2 & \cdots & d_n \\ 0 & d_0 & d_1 & \cdots & d_{n-1} \end{bmatrix} \mathbf{a}(\mathbf{q}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (35)$$

Numerical values for α may be substituted into (34) and (35). Alternatively the expression may first be simplified symbolically. Premultiplying

(35) by

$$\mathbf{S} = \begin{bmatrix} 0 & 1 \\ -1 & 2\alpha \end{bmatrix} \quad (36)$$

yields

$$\begin{bmatrix} 0 & d_0 & d_1 & d_2 & \cdots & d_{n-1} \\ -d_0 & 0 & d_0 & d_1 & \cdots & d_{n-2} \end{bmatrix} \mathbf{a}(\mathbf{q}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (37)$$

The term d_n with the highest power of α has been removed. This reduction procedure may be continued $n/2$ times for n even and $(n+1)/2$ times for n odd. The resulting equations are for n even

$$\begin{bmatrix} -d_{n/2-2} & \cdots & -d_0 & 0 & d_0 & d_1 & \cdots & d_{n/2} \\ -d_{n/2-1} & \cdots & -d_1 & -d_0 & 0 & d_0 & \cdots & d_{n/2-1} \end{bmatrix} \mathbf{a}(\mathbf{q}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (38)$$

and for n odd

$$\begin{bmatrix} -d_{(n+1)/2-2} & \cdots & -d_0 & 0 & d_0 & d_1 & \cdots & d_{(n+1)/2-1} \\ -d_{(n+1)/2-1} & \cdots & -d_1 & -d_0 & 0 & d_0 & \cdots & d_{(n+1)/2-2} \end{bmatrix} \mathbf{a}(\mathbf{q}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (39)$$

In (38) and (39) the highest degree polynomials in α have been reduced to about half the degree arising in (35).

Example 3. Logarithmic spiral $z = e^{i\alpha} \cdot e^{i\eta}$. This curve is of interest for sampled-data systems. Constant damping lines $\sigma = \pm c\omega$ ($c = -\zeta/\sqrt{1-\zeta^2}$, ζ = damping) are mapped via $z = e^{sT} = e^{\sigma T} \cdot e^{j\omega T}$ into the z -plane. Let $\alpha = \omega T$, $\alpha \in [-\pi; 0]$, $z = r + j\eta = e^{i\alpha} \cos \alpha + je^{i\alpha} \sin \alpha$. τ and η now play the role of σ and ω in (24)

$$\begin{aligned} d_0 &= 1 \\ d_1 &= 2\tau = 2e^{i\alpha} \cos \alpha \\ d_{i+1} &= 2\tau d_i - (\tau^2 + \omega^2) d_{i-1} \\ &= 2e^{i\alpha} \cos \alpha d_i - e^{2i\alpha} d_{i-1}. \end{aligned} \quad (40)$$

Remark 1. In (36) a simplification was introduced that reduces the degree of the polynomials in α for the example. The question may arise, for what other examples this idea may be useful. It is easily verified that k premultiplications of (23) by

$$\mathbf{S} = \begin{bmatrix} 0 & \sigma^2(\alpha) + \omega^2(\alpha) \\ -1 & 2\sigma(\alpha) \end{bmatrix} \quad (41)$$

modify it to

$$[\sigma^2 + \omega^2]^k \begin{bmatrix} c_{k-2} & c_{k-3} & \cdots & c_0 & 0 & d_0 & d_1 & \cdots & d_{n-k} \\ c_{k-1} & c_{k-2} & \cdots & c_1 & c_0 & 0 & d_0 & \cdots & d_{n-k-1} \end{bmatrix} \mathbf{a}(\mathbf{q}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (42)$$

where $c_i = -d_i/[\sigma^2 + \omega^2]^{i+1}$. A simplification in terms of α is only obtained if $\sigma^2(\alpha) + \omega^2(\alpha)$ is a simple expression, e.g. R^2 for a circle with center zero and radius R . For Example 3 the reduction matrix is

$$S = \begin{bmatrix} 0 & e^{2\alpha} \\ -1 & 2e^{\alpha} \cos \alpha \end{bmatrix}. \quad (43)$$

For boundaries with more complicated functions $\sigma^2(\alpha) + \omega^2(\alpha)$ it is recommended to use the original form (23).

It should be noted however, that the resulting equations in \mathbf{q} can be modified by the above premultiplication. If, for example, only one coefficient $a_i(\mathbf{q})$ is a nasty nonlinear function of q_1 and all other coefficient functions are linear in q_1 , then a form (42) can be chosen in which $a_i(\mathbf{q})$ is multiplied by zero in one of the equations and it can be solved for q_1 .

Remark 2. Note that a related boundary representation was derived by Šiljak (1969). Instead of our starting point (25) he uses the real and imaginary part of $P(s)$ where $s^k = X_k + jY_k$. Then

$$\begin{aligned} P(s) &= \operatorname{Re} P(X_1 + jY_1) + j \operatorname{Im} P(X_1 + jY_1) \\ &= \sum_{k=0}^n a_k X_k + j \sum_{k=0}^n a_k Y_k = 0. \end{aligned} \quad (44)$$

the two equations for real and imaginary parts are related to (23) by

$$\begin{bmatrix} \operatorname{Re} P \\ \operatorname{Im} P \end{bmatrix} = \begin{bmatrix} 1 & -\sigma(\alpha) \\ 0 & \omega(\alpha) \end{bmatrix} \begin{bmatrix} d_0(\alpha) & d_1(\alpha) & \cdots & d_n(\alpha) \\ 0 & d_0(\alpha) & \cdots & d_{n-1}(\alpha) \end{bmatrix} \mathbf{a}. \quad (45)$$

The advantage of the form (23) is that only one recursion for the d_k is needed instead of two recursions for X_k and Y_k .

Remark 3. Equation (45) is useful also for the zero exclusion approach to robustness analysis [see Barmish (1988)]. The robustness test is based on the condition

$$0 \notin \{P(s, \mathbf{q}) : \mathbf{q} \in Q, s \in \partial\Gamma\} \quad (46)$$

For the complex boundary, $\omega \neq 0$, this is equivalent to

$$0 \notin \{X(\alpha, \mathbf{q}) + jY(\alpha, \mathbf{q}) : \mathbf{q} \in Q, \alpha \in I_\alpha\} \quad (47)$$

where

$$\begin{bmatrix} X(\alpha, \mathbf{q}) \\ Y(\alpha, \mathbf{q}) \end{bmatrix} = \begin{bmatrix} d_0(\alpha) & d_1(\alpha) & \cdots & d_n(\alpha) \\ 0 & d_0(\alpha) & \cdots & d_{n-1}(\alpha) \end{bmatrix} \mathbf{a}(\mathbf{q}) \quad (48)$$

All calculations are done in the real field

4. SOLUTION OF THE BOUNDARY EQUATIONS FOR q_1 AND q_2

For a given polynomial

$$P(s, \mathbf{q}) = [1, s, s^2, \dots, s^n] \mathbf{a}(\mathbf{q}) \quad (49)$$

the complex root boundary in \mathbf{q} -space is given by the \mathbf{q} -values satisfying (23). As in the example of Section 2 an $l=2$ dimensional grid for $\mathbf{q}_0 = [q_1, q_2, \dots, q_l]^T$ in Q must be chosen. For each grid point $\mathbf{q}_0 = \mathbf{q}_0^*$ (23) may be solved for $q_1(\mathbf{q}_0^*, \alpha)$ and $q_2(\mathbf{q}_0^*, \alpha)$. For robust stability q_1 and q_2 must be outside the interval (5) for all $\alpha \in I_\alpha$. This condition may be tested graphically like in Fig. 1.

The difficulty of solving (23) for q_1 and q_2 obviously depends on the kind of coefficient functions $\mathbf{a}(\mathbf{q})$. In the example (16) the solution of the two equations (17) for q_1 and q_2 is easy, because the coefficient functions $\mathbf{a}(\mathbf{q})$ are affine linear in q_1 and q_2 . In this section the solution for polynomial coefficient functions in q_1 and q_2 is given. In this case (23) may be written

$$\begin{aligned} f &= f_0(q_2) + f_1(q_2)q_1 + \cdots + f_k(q_2)q_1^k = 0 \\ g &= g_0(q_2) + g_1(q_2)q_1 + \cdots + g_m(q_2)q_1^m = 0 \end{aligned} \quad (50)$$

The f_k and g_k are polynomials in q_2 and continuous functions of \mathbf{q}_0 and α . k and m are chosen such that f_k and g_m are not identically zero.

For each \mathbf{q}_0 and α , (50) describes the intersections of two algebraic curves in the (q_1, q_2) -plane. These solutions q_1, q_2 of (50) are calculated by elimination of one variable, say q_1 , by the resultant method, finding the real roots $q_2^{(r)}$. Then the corresponding $q_1^{(r)}$ are given by a linear equation (in nonsingular cases). The resultant method is also contained in software packages for symbolic manipulations like REDUCT, [see Hearn (1987)].

The method of finding *all* real intersection points of (50) is reviewed here following Brill (1925).

Expand (50) by the redundant equations

$$\begin{aligned} f &= 0 & g &= 0 \\ fq_1 &= 0 & gq_1 &= 0 \\ &\vdots & &\vdots \\ fq_1^{m-1} &= 0 & gq_1^{k-1} &= 0 \end{aligned} \quad (51)$$

and write in matrix notation

$$\begin{bmatrix} f \\ \vdots \\ fq_1^{m-1} \\ g \\ \vdots \\ gq_1^{k-1} \end{bmatrix} = \begin{bmatrix} f_0 & f_1 & & f_k & 0 & 1 \\ 0 & f_0 & & & & q_1 \\ & & & & & q_1^2 \\ 0 & & f_0 & f_1 & f_k & \\ g_0 & g_1 & & g_m & 0 & \\ 0 & g_0 & & & & \\ & & & & & \\ 0 & & g_0 & g_1 & & g_m & q_1^{k+m-1} \end{bmatrix} \tag{52}$$

The vector $[1, q_1, \dots, q_1^{k+m-1}]$ is nonzero, thus (52) is equivalent to the condition $R = \det \mathbf{R} = 0$. The resultant R is a polynomial in q_2 . It is easily generated symbolically for general q_2 , \mathbf{q}_N and α , e.g. by a procedure provided by REDUCE. For finding the (real) zeros we have to substitute numerical values for the grid points \mathbf{q}_N^* and α^* . Let $q_2 = q_2^{(i)}$ be a real root of $R(q_2) = 0$.
Now determine the corresponding real value $q_1^{(i)}$, that satisfies both equations (50) with $q_2 =$

$q_2^{(i)}$ simultaneously. q_1 may be determined from (52). Note that in the generic case rank $\mathbf{R} = k + m - 1$, thus one linear dependent row and column can be removed from (52). A further reduction of the matrix dimension by one to a square of dimension $k + m - 2$ can be achieved by allowing q_1 to appear also in the matrix. Both steps can be accomplished by removing the equations $fq_1^{m-1} = 0$ and $gq_1^{k-1} = 0$ from (52).

$$\begin{bmatrix} f & f_0 & f_1 & & f_k & 0 \\ & & & & & \\ fq_1^n & 0 & & f_0 & f_1 & f_k \\ & g_0 & g_1 & & g_m & 0 \\ & & & & & \\ gq_1^k & 0 & & g_0 & g_1 & g_m \end{bmatrix} \begin{bmatrix} 1 \\ q_1 \\ q_1^2 \\ \vdots \\ \vdots \\ q_1^{k+m-2} \end{bmatrix} \tag{53}$$

This rectangular matrix is now made square by including the last column into the last but one as

$$\begin{bmatrix} f_0 & & f_{k-1} & f_k & & 0 \\ & & & & & \\ & & & & f_{k-1} & f_k \\ 0 & & f_0 & & f_{k-2} & f_{k-1} + f_k q_1 \\ g & g_0 & & g_{m-1} & g_m & 0 \\ & & & & & \\ & & & & g_{m-1} & g_m \\ gq_1^k & 0 & & g_0 & g_{m-2} & g_{m-1} + g_m q_1 \end{bmatrix} \begin{bmatrix} 1 \\ q_1 \\ q_1^2 \\ \vdots \\ \vdots \\ q_1^{k+m} \end{bmatrix} = \mathbf{0} \tag{54}$$

(54) implies $T = \det \mathbf{T} = 0$, and q_1 can be calculated as the ratio of two determinants

$$T = \begin{bmatrix} f_0 & & f_{k-1} & f_k & & 0 \\ & & & & & \\ & & & & & f_k \\ 0 & f_0 & & f_{k-2} & f_{k-1} & \\ g_0 & & g_{m-1} & g_m & & \\ & & & & & g_m \\ 0 & g_0 & & & & g_{m-1} \end{bmatrix} + q_1 \begin{bmatrix} f_0 & & f_{k-1} & f_k & 0 & 0 \\ & & & & f_k & 0 \\ & & & & f_{k-1} & 0 \\ 0 & f_0 & & f_{k-2} & f_{k-1} & f_k \\ g_0 & & g_{m-1} & g_m & 0 & 0 \\ & & & & g_m & \\ & & & & g_{m-1} & 0 \\ 0 & g_0 & & & g_{m-2} & g_m \end{bmatrix} = 0 \quad (55)$$

T_1 T_2

$q_1 = -\det T_1 : \det T_2$ (56)

If $\det T_2 = 0$ and $\det T_1 \neq 0$, then q_1 goes to infinity for that q_2 . If both determinants vanish, then the algebraic curve is reducible.

A singular case arises, if in (52) rank $R < k + m - 1$. In a similar way we remove additional columns and include more columns to square the matrix. Again we arrive at $\det T = 0$, but now two entries in T are of quadratic, third order etc. depending on the rank deficiency of R [see Brill (1925)]. In this case two or more values of q_1 correspond to one value of q_2 .

Remark 4. The reader may wonder why we dig out an old elimination method instead of using the modern approach of Groebner bases [see Buchberger (1985)], that is also contained in REDUCE 3.3. The method of Groebner bases does a symbolic evaluation of $\det T_1$ and $\det T_2$ for arbitrary α and q_2 and yields very complicated expressions. For their numerical evaluation, however, $q_2 = q_2(\alpha)$ must be calculated by factorizing the polynomial $R(\alpha, q_2)$ for given α anyway. In the example of the next section it was not possible to compute a Groebner basis in reasonable time.

5. APPLICATION TO A TRACK-GUIDED BUS

A Daimler Benz 0305 bus is guided by the field generated by a wire in the street (Darenberg, 1986). The linearized system with actuator input u = steering angle rate, and output y = displacement of front antenna from the guideline, has the transfer function

$$G(s, q_1, q_2) = \frac{q_1(48280q_1 + 388600s + 609.8q_1q_2s^2)}{(16.8q_1^2q_2 + 270000) + 1077q_1q_2s + q_1^2q_2^2s^2} s^4$$

(57)

Note that initially the problem had four interval parameters:

- m mass
- J moment of inertia with respect to vertical axis

- μ friction coefficient (1 for dry road, 0.5 for wet road)
- v velocity.

There is a relationship $J = cm$ with a constant c for the empty bus and for the full bus. In between these cases the passengers could cluster either in the center or at the ends of the bus and thereby change c . This second-order effect was neglected by the psychological assumption that passengers distribute uniformly over the bus. After substitution of $J = cm$ it turned out that m and μ did not appear separately in the dynamic equations but only as one parameter $q_2 = m/\mu$, $q_2 \in [9.95 \text{ tons}, 32 \text{ tons}]$. The first parameter is the bus velocity $q_1 = v$. For $v = 0$ the bus is not controllable and $G(s, 0, q_2) = 0$. Γ -stability is required in the speed range $q_1 \in [3 \text{ ms}^{-1}, 20 \text{ ms}^{-1}]$.

The resulting operating domain is shown in Fig. 3. The desired Γ region is the left side of a hyperbola, i.e. points $s = \sigma + j\omega$ satisfying

$$\left(\frac{\sigma}{0.35}\right)^2 + \left(\frac{\omega}{1.75}\right)^2 = 1, \quad \sigma = \alpha,$$

$\alpha \in (-\infty; -0.35].$ (58)

Since this paper is concerned with analysis the controller $C(s)$ is taken as given.

$$C(s) = \frac{9375 + 10938s + 2344s^2}{15625 + 1250s + 50s^2 + s^4}$$

(59)

$C(s)$ was determined by Muench (1986) using

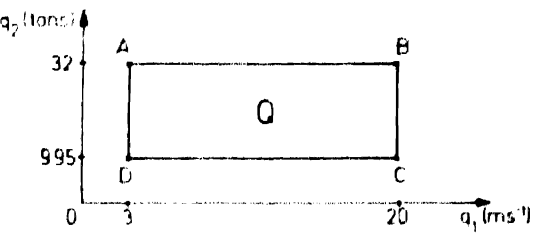


FIG. 3. Operating domain of the automatic bus steering feedback system. q_1 = velocity, $q_2 = m/\mu$, where m = mass, μ = friction coefficient.

simultaneous pole region assignment for the representative operating conditions A, B, C and D in Fig. 3. Thus there exists a $\mathbf{q}_0 \in Q$ such that the closed loop is Γ -stable and condition (i) of Frazer and Duncan (1929) is satisfied.

The closed-loop characteristic polynomial is

$$\begin{aligned} P(s, q_1, q_2) &= \text{numerator of } [1 + C(s)G(s, q_1, q_2)] \\ &= \sum_{i=0}^n a_i(q_1, q_2)s^i. \end{aligned}$$

A lengthy but straightforward computation gives the coefficients

$$\begin{aligned} a_0 &= 453 \times 10^6 q_1^2 \\ a_1 &= 528 \times 10^6 q_1^2 + 3640 \times 10^6 q_1 \\ a_2 &= 5.72 \times 10^6 q_1^2 q_2 + 113 \times 10^6 q_1^2 + 4250 \times 10^6 q_1 \\ a_3 &= 6.93 \times 10^6 q_1^2 q_2 + 911 \times 10^6 q_1 + 4220 \times 10^6 \\ a_4 &= 1.45 \times 10^6 q_1^2 q_2 + 16.8 \times 10^6 q_1 q_2 + 338 \times 10^6 \\ a_5 &= 15.6 \times 10^3 q_1^2 q_2^2 + 840 q_1^2 q_2 + 1.35 \\ &\quad \times 10^6 q_1 q_2 + 13.5 \times 10^6 \\ a_6 &= 1.25 \times 10^3 q_1^2 q_2^2 + 16.8 q_1^2 q_2 + 53.9 \\ &\quad \times 10^3 q_1 q_2 + 270 \times 10^3 \\ a_7 &= 50 q_1^2 q_2^2 + 1080 q_1 q_2 \\ a_8 &= q_1^2 q_2^2. \end{aligned} \quad (60)$$

The real root boundary for $\sigma = -0.35$ is described by

$$\begin{aligned} P(-0.35, q_1, q_2) &= -79.66(q_1^2 q_2^2 - 5339 q_1^2 q_2 \\ &\quad - 3077 q_1 q_2 - 3540213 q_1^2 \\ &\quad + 9946676 q_1 + 2208293) = 0. \end{aligned} \quad (61)$$

The complex root boundary $\partial\Gamma$ is parameterized by $\sigma(\alpha) = \alpha$, $\omega^2(\alpha) = 25\alpha^2 - 1.75^2$, $\alpha \in (-\infty; -0.35]$. Hence any real polynomial which vanishes at $s(\alpha) = \sigma(\alpha) + j\omega(\alpha)$ must have the factor

$$\begin{aligned} \sigma^2(\alpha) + \omega^2(\alpha) - 2\sigma(\alpha)s + s^2 \\ = (26\alpha^2 - 3.0625) - 2\alpha s + s^2 \end{aligned}$$

and the coefficients in the boundary representation (23) are given by

$$\begin{aligned} d_0(\alpha) &= 1 \\ d_1(\alpha) &= 2\alpha \end{aligned} \quad (62)$$

$$d_{i+1}(\alpha) = 2\alpha d_i(\alpha) - (26\alpha^2 - 3.0625)d_{i-1}(\alpha).$$

Substituting the d_i and $\mathbf{a}(\mathbf{q})$ into (23) and collecting terms with the same power of q_1 the following form is obtained:

$$\begin{aligned} f_0(\alpha) + [f_{10}(\alpha) + f_{11}(\alpha)q_2]q_1 \\ + [f_{20}(\alpha) + f_{21}(\alpha)q_2 + f_{22}(\alpha)r_2^2]r_1^2 = 0 \end{aligned} \quad (63)$$

$$\begin{aligned} g_0(\alpha) + [g_{10}(\alpha) + g_{11}(\alpha)q_2]q_1 \\ + [g_{20}(\alpha) + g_{21}(\alpha)q_2 + g_{22}(\alpha)q_2^2]q_1^2 = 0. \end{aligned} \quad (64)$$

The resultant of (63) and (64) has the form

$$\begin{aligned} R(q_2) &= h_0(\alpha) + h_1(\alpha)q_2 + h_2(\alpha)q_2^2 \\ &\quad + h_3(\alpha)q_2^3 + h_4(\alpha)q_2^4 = 0. \end{aligned} \quad (65)$$

If all its roots are complex for a given $\alpha = \alpha^*$, then there exists no real pair q_1, q_2 for which the closed loop has an eigenvalue at $\sigma(\alpha^*) + j\omega(\alpha^*)$. On the other hand if there are real solutions $q_2^{(i)}(\alpha^*)$, then the corresponding $q_1^{(i)}(\alpha^*)$ is given by the root of the greatest common divisor of (63) and (64). By (55) and (56) for

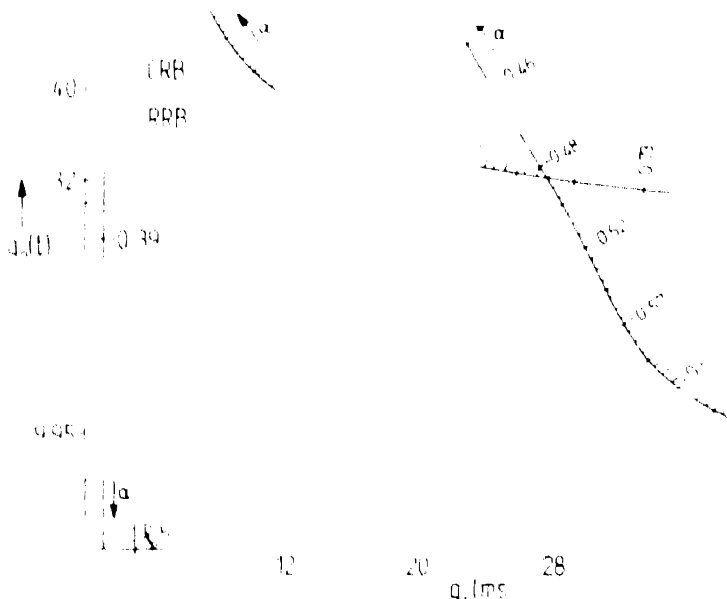


FIG. 4. The track-guided bus is robustly Γ -stable [see (58)] in the operating domain Q . For a speed of 3 m s^{-1} the real root at $\sigma = -0.35$ is critical. α is the real part of the eigenvalue on the complex boundary. For 20 m s^{-1} $\alpha_1 = -0.69$ is the critical pole location for the full bus and $\alpha_2 = -2.97$ for the empty bus.

each fixed $\alpha = \alpha^*$ and $q_2 = q_2(\alpha^*)$

$$q_1 = \begin{bmatrix} f_0 & f_1 \\ g_0 & g_1 \end{bmatrix} : \begin{bmatrix} f_0 & f_2 \\ g_0 & g_2 \end{bmatrix} \\ \begin{bmatrix} f_0 & f_{10} + f_{11}q_2 \\ g_0 & g_{10} + g_{11}q_2 \end{bmatrix} : \begin{bmatrix} f_0 & f_{20} + f_{21}q_2 + f_{22}q_2^2 \\ g_0 & g_{20} + g_{21}q_2 + g_{22}q_2^2 \end{bmatrix} \quad (66)$$

For physical reasons we are only interested in positive solutions $q_1 > 0$, $q_2 > 0$. Figure 4 shows the Γ -stability boundaries. The vertical line at $v = 3 \text{ ms}^{-1}$ is the real root boundary (RRB). To the left of it is a branch of the complex root boundary (CRB), further branches are the curves to the right. It is seen that the operating domain Q is Γ -stable.

In fact we can read from Fig. 4, how much Q could be expanded without losing Γ -stability, e.g. the velocity may be increased to 27.5 ms^{-1} , if all other circumstances allow a safe operation at this speed. For the maximum design speed of 20 ms^{-1} the critical generalized frequencies at the closest boundary points are $\alpha_1 = -0.69$ (i.e. $s(\alpha_1) = -0.69 \pm j2.97$) for the full bus and $\alpha_2 = -2.97$ (i.e. $s(\alpha_2) = -2.97 \pm j14.75$) for the empty bus. At the minimum speed of 3 ms^{-1} the eigenvalue at $s = -0.35$ is the critical one.

6. CONCLUSIONS

A parameter space method for robustness analysis was derived. It is based on a boundary representation theorem for the complex root boundary. Only continuity of the coefficients of the closed-loop characteristic polynomial with respect to the uncertain parameters must be assumed; otherwise the boundary representation is completely general and gives necessary and sufficient conditions for Γ -stability of a given operating domain. A broad class of regions Γ for the desired eigenvalue locations can be treated. The complex boundary is described by two equations. It depends on the kind of coefficient functions, how easy or how difficult it is to reduce the two equations to one by eliminating one of the uncertain parameters. For the cases of affine, multilinear and polynomial coefficient functions the general solution is shown. Eventually $l - 2$ of the l uncertain parameters must be gridded. The final numerical calculation and test

for each grid point may be done numerically. However additional insight beyond a yes-no answer to the Γ -stability question can be gained by a graphical illustration of the computed boundaries in projection on a two dimensional subspace of the parameter space.

The method is recommended for problems with a small number of uncertain parameters and polynomial coefficient functions. One example of this kind is a track-guided bus, for which a robustness analysis is performed.

REFERENCES

- Ackermann, J. and B. R. Barmish, (1988) Robust Schur stability of a polytope of polynomials. *IEEE Trans. Aut. Control*, **33**, 984-986.
- Ackermann, J., H. Z. Hu and D. Kaesbauer (1988) Robustness analysis: a case study. *Preprints CDC*, Austin, TX, **1**, 86-91.
- Ackermann, J. and D. Kaesbauer (1981) D-decomposition in the space of feedback gains for arbitrary pole regions. *Preprints 8th IFAC Congress*, Kyoto, **4**, 12-17.
- Bartlett, A. C., C. V. Hollot and Huang Lin (1988) Root location of an entire polytope of polynomials: It suffices to check the edges. *J. Math. Control Signals Syst.*, **1**, 61-71.
- Barmish, B. R. (1988) New Tools for robustness analysis. *Proc. 27th Conf. on Decision and Control*, Austin, TX, **1**, 1-6.
- Bialas, S. (1985) A necessary and sufficient condition for the stability of convex combinations of stable polynomials or matrices. *Bull. Polish Acad. Sci., Tech. Sci.*, **33**, 473-480.
- Brill, A. (1925) *Vorlesungen ueber ebene algebraische Kurven und algebraische Funktionen*. Braunschweig, Vieweg.
- Buchberger, B. (1985) Groebner bases, An algorithmic method in polynomial ideal theory. In N. K. Bose (Ed.), *Multidimensional Systems Theory*, pp 184-232. Reidel, Hingham, MA.
- Darenberg, W. (1986) Automatische Spurfuehrung von Kraftfahrzeugen. Ein Problem der robusten Regelung. *GMA-Bericht*, **11**, Aussprachetag Robustic Regelung, Langen, pp 45-64.
- Frazer, R. A. and W. J. Duncan (1929) On the criteria for the stability of small motions. *Proc. R. Soc. A*, **125**, 642-654.
- Gantmacher, F. R. (1959) *The Theory of Matrices*. Chelsea, New York.
- Hearn, A. C. (1987) *REDUCE User's Manual*. The RAND Corporation, Santa Monica, CA.
- Khantonov, V. I. (1978) Asymptotic stability of an equilibrium position of a family of systems of linear differential equations. *Differentsial'nye Uravneniya*, vol. **14**, 2086-2088. [Engl. Transl. (1979) *Differential Equations*. Plenum Press, New York.]
- Muench, R. (1986) Erweiterung des Parameterraumverfahrens und Anwendung auf einen spurfuehrenden Bus. Diplomarbeit TU Muenchen.
- Šiljak, D. D. (1969) *Nonlinear Systems*. John Wiley, New York.
- Zadeh, L. A. and C. A. Desoer (1963) *Linear System Theory—the State Space Approach*. McGraw-Hill, New York.

On the Attitude Stabilization of Rigid Spacecraft*

CHRISTOPHER I. BYRNES† and ALBERTO ISIDORI‡

While rigid body models for spacecraft with two controls are locally controllable and locally reachable for most actuator configurations, these systems cannot be locally asymptotically stabilized by smooth feedback, but using methods from a general nonlinear feedback design theory, feedback laws are derived which control the closed-loop trajectories to a revolute motion about an axis of rotation.

Key Words—Attitude control, stability, nonlinear control systems, nonlinear systems, feedback control.

Abstract—In this paper, we settle in the negative a longstanding problem concerning the existence of a smooth (static or dynamic) state variable feedback law locally asymptotically stabilizing a rigid spacecraft with two controls about a desired reference attitude. Modelling a spacecraft actuated by three thruster jets, one of which has failed, this well studied system is known to be locally reachable and locally asymptotically null controllable. We obtain our result as a corollary of a surprising result which asserts, for a class of nonlinear systems containing several examples of interest, that such a system is locally asymptotically stabilizable precisely when it can be linearized via state feedback transformations. We give a further result on the instability (in the sense of Lyapunov) of rigid spacecraft for certain feedback laws, but we are able to construct a feedback law locally asymptotically driving the closed-loop trajectories to a motion about the third principal axis. This law is derived using general principles comprising a nonlinear enhancement of root-locus design principles.

1. INTRODUCTION

IN THIS PAPER, we study the basic problem of constructing smooth state variable feedback laws achieving certain desired steady-state closed-loop behaviour for a system, essentially Euler's equations for a rigid body, modelling a rigid spacecraft actuated by three thruster jets, one of

which is in a failure mode. This problem has been extensively studied in the literature [see e.g. Aeyels (1985), Brockett (1983), Crouch (1984, 1985), Hermes (1980)]. For example, it is known that, relative to a fixed desired reference attitude, this system is locally reachable and locally asymptotically null controllable. It has been a longstanding problem to realize such open-loop strategies via feedback, e.g. to find a smooth (e.g. C^1) state feedback law locally asymptotically stabilizing this system. In Section 2 we prove a general stability result valid for a class of nonlinear systems containing several examples of interest, specifically the rigid spacecraft model in question. Perhaps surprisingly, a corollary of our main result asserts that there exists a smooth feedback law locally asymptotically stabilizing such a system if, and only if, the system can be linearized via coordinate changes in the state and input spaces and state feedback transformations. As a further corollary, we show that there exists no smooth (static or dynamic) state feedback law locally stabilizing this rigid spacecraft system about a nominal reference frame. While examples of locally reachable but not static feedback stabilizable systems have been recorded earlier [Brockett (1983); Aeyels (1985); Sontag and Sussmann (1980)], we stress that this particular system is a widely used model for a physical system, demonstrating that the nonexistence of a smooth closed-loop realization of an open-loop control is not just a mathematical anomaly.

The proof (Section 3) entails on the one hand an application of well-known linearization

* Received 14 October 1987, revised 17 April 1989, received in final form 8 May 1989. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor V. Utkin under the direction of Editor H. Kwakernak.

† Department of Systems Science and Mathematics, Washington University, St. Louis, MO 63130, U.S.A.

‡ Dipartimento di Informatica e Sistemistica, Università di Roma, Roma, Italia.

techniques [see e.g. Brockett (1978); Jakubczyk-Respondek (1980); Sommer (1980); Hunt *et al.* (1983)] to this class of systems. We remark that for mechanical systems, our proof boils down to the existence of an independent control for every position variable. It is more likely this feature, rather than including in the system design specific involutivity considerations, which accounts for the relatively ubiquitous application of linearization techniques. On the other hand, the converse of our theorem is proved by appealing to Krasnoselski's Theorem which computes the degree of a locally asymptotically stable vector field. As pointed out by Zabczyk (1989), this degree calculation also implies a local solvability criterion discovered independently by Brockett (1983). Moreover, from Zabczyk's argument it would appear that the nonexistence of smooth stabilizing laws for this satellite model also applies to Lipschitz continuous feedback laws.

In Section 4, we briefly illustrate our general methods for applying center manifold theory to state feedback design, based on a nonlinear enhancement of classical root-locus design principles. We apply these methods in Section 5 to derive a state feedback law which (locally) asymptotically drives the closed-loop trajectories of the satellite to a motion about the third principal axis, thereby stabilizing the motion about an attractor, which is a circle and not an isolated equilibrium. This control scheme raises several interesting questions about the steady-state limits of such trajectories; e.g. does this control scheme, or does any continuous feedback law, render the satellite dynamics Lyapunov stable? For example, the open-loop dynamics are Poisson stable and therefore not Lyapunov stable. We conclude with a fairly general instability result, asserting in particular that robustly (Lyapunov) stabilizing feedback laws asymptotically inducing a motion about the third principal axis do not exist.

2 ASYMPTOTIC STABILIZABILITY OF A CLASS OF NONLINEAR SYSTEMS

In this section we shall investigate the local asymptotic stabilizability of control systems evolving on $R^{n_1+n_2}$ having the form

$$\dot{x}_1 = f_1(x_1, x_2)x_1 + \sum_{i=1}^m b_i u_i, \quad u_i \in R, \quad x_1, b_i \in R^n \quad (2.1a)$$

$$\dot{x}_2 = f_2(x_1, x_2), \quad x_2 \in R^{n_2} \quad (2.1b)$$

where in addition we shall assume:

(H1) The drift vector field

$$f(x) = \begin{pmatrix} f_1(x)x_1 \\ f_2(x) \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in R^{n_1+n_2}$$

is C^1 and has 0 as an equilibrium, i.e. $f(0) = 0$;

(H2) $f_2(x_1, x_2) = 0$ implies $x_1 = 0$; and

(H3) The Jacobian matrix $\partial f_2 / \partial x_1(0)$ has rank n_2 .

Examples of systems (2.1) satisfying (H1)–(H3) abound, since many controlled “second-order” systems take the form (2.1) when expressed as a “first-order” system. In particular, the rigid body model for an n -joint robot arm, operating in an environment where gravitational forces have been cancelled or are negligible (such as outer space or underwater applications) and controlled at each joint by an applied torque (e.g. actuated by a DC motor)

$$M(q)\ddot{q} + B(q, \dot{q})\dot{q} = \tau, \quad q \in R^n \quad (2.2)$$

can be expressed in the form (2.1) after a smooth, state-dependent change of coordinates in input space with state

$$x = \begin{pmatrix} q \\ \dot{q} \end{pmatrix} \in R^{2n}. \quad (2.3)$$

Of particular interest in the present paper is the rigid body model of a satellite controlled by momentum exchange devices, such as momentum wheels or gas jet actuators. On the state manifold $M = SO(3) \times R^3$, the evolution of an orientation, angular velocity pair (R, ω) takes the form

$$J\dot{\omega} = S(\omega)J\omega + \sum_{i=1}^m b_i u_i \quad (2.4a)$$

$$\dot{R} = S(\omega)R \quad (2.4b)$$

where J is the inertia matrix and $S(\omega)$ is the matrix representation of the cross-product, $b \rightarrow b \times \omega$; i.e.

$$S(\omega) = \begin{pmatrix} 0 & \omega_3 & -\omega_2 \\ -\omega_3 & 0 & \omega_1 \\ \omega_2 & -\omega_1 & 0 \end{pmatrix} \quad (2.5)$$

Choosing principle axes (i.e. diagonalizing J), (2.4)(2.5) can be expressed in local coordinates about a reference frame $R = [r_1 r_2 r_3]$ using Euler angles, φ, θ, ψ representing rotations about the r_1, r_2, r_3 axes, respectively. Explicitly (2.4) takes

the form

$$\begin{bmatrix} \dot{\omega}_1 \\ \dot{\omega}_2 \\ \dot{\omega}_3 \end{bmatrix} = \begin{bmatrix} a_1 \omega_2 \omega_3 \\ a_2 \omega_1 \omega_3 \\ a_3 \omega_1 \omega_2 \end{bmatrix} + \sum_{i=1}^m b_i u_i, \quad b_i \in R^3 \quad (2.4a)$$

$$\begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ \sin \theta \tan \phi & 1 & -\cos \theta \tan \phi \\ -\sin \theta \sec \phi & 0 & \cos \theta \sec \phi \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} \quad (2.4b)'$$

In particular, (2.4)' has the form (2.1) with $n_1 = n_2 = 3$ and satisfies (H1)–(H2) and (H3), since

$$\det \frac{\partial f_2}{\partial x_1}(0) = I.$$

In general, we shall set

$$m' = \dim \text{span} \{b_1, \dots, b_m\}$$

in (2.1).

The local control of the satellite equations (2.4)' is well-known to be relatively straightforward if $m' \geq 3$, but the case $m' = 2$ is far more difficult. While it is known in this case that the satellite equations are locally reachable and locally null controllable, it has been a longstanding, open problem as to whether (2.4)' can be controlled to 0 using a closed-loop strategy, especially using continuously differentiable state feedback. One of the corollaries to our first set of theorems is that this widely believed conjecture is false.

Theorem 1. Consider a system (2.1) satisfying (H1)–(H3). There is a continuously differentiable feedback law, $u_i = F_i(x)$, rendering the origin locally asymptotically stable if, and only if, $m' = n_1$.

Necessity is a consequence of our next theorem which will be demonstrated in Section 3.

Theorem 2. Consider a system (2.1) satisfying (H1)–(H2). If there is a continuously differentiable static or dynamic state feedback law rendering the origin locally asymptotically stable, then $m' = n_1$.

Corollary 1. A rigid satellite with (one or) two independent actuators cannot be locally asymptotically stabilized using continuously differentiable static or dynamic state feedback.

While several examples of locally reachable or locally asymptotically null controllable systems

which cannot be stabilized have been known for some time we stress that this particular system is a widely used model for a physical system, not a mathematical anomaly. Concerning the sufficiency of the condition $m' = n_1$, we first remark that for a rigid robot arm (2.2), we have $m = m' = n$ and linearization, known in this case as feedback stabilization by the method of "computed torque", has been widely used for some time. More generally, if $m' = n_1$, $\text{span} \{g_1, \dots, g_m\}$ is trivially involutive (where $g_i = (b_i^T 0^T)^T$) and by (H3) we have

$$\begin{aligned} \text{span} \{g_1, \dots, g_m, \text{ad}_f(g_1)(0), \dots, \text{ad}_f(g_m)(0)\} \\ = T_0(R^{n_1+n_2}). \end{aligned}$$

In particular, if $m' = n_1$ then (2.1) can be linearized using state and input coordinate changes and state feedback, from which sufficiency follows [see e.g. Jakubczyk and Respondek (1980)]. Therefore, a nonlinear system (2.1) satisfying (H1)–(H3) can be stabilized by smooth feedback if, and only if, it can be linearized. Of course, systems (2.1) such as (2.2) or (2.4) are not designed so that certain distributions are involutive. Rather, they are designed so that there is an independent control for virtually everything in the system which moves, i.e. so that $m' = n_1$. For mechanical systems of this form, taking position coordinates as outputs this amounts to designing the system to be "square". Moreover, the conditions (H1), (H3) and $m' = n_1$ imply the condition that about the equilibrium 0, the linearization of (2.1) is controllable. It is for these reasons that linearization techniques apply so generally within the class of systems (2.1). We summarize these results as follows:

Theorem 3. Consider a system (2.1) satisfying (H1)–(H3). The following are equivalent:

- (1) The system can be locally asymptotically stabilized by smooth state feedback.
- (2) The system can be locally asymptotically stabilized by smooth dynamic state feedback.
- (3) $m' = n_1$.
- (4) The system can be linearized by smooth coordinate changes and smooth feedback.
- (5) The linearization (i.e. the first variation) about 0 is controllable.

Thus, a rigid satellite can be locally asymptotically stabilized about a nominal reference attitude when we retain full control, i.e. when $m' = 3$. Nonetheless, some stabilization results in the case $m' = 2$ were obtained by several authors for certain subsets of the controlled rigid body equations (2.4a)'–(2.4b)'. For example, Brock-

ett (1983), Aeyels (1985), Crouch and Irving (1983) and Byrnes and Isidori (1988a) derive stabilization schemes for the three angular momentum equations (see also Section 5). Assuming the two actuators are aligned along the first two principal axes, Hermes (1980) derived stabilizing laws for the four equations involving ω_1 , ω_2 , ϕ , θ so as to induce a spinning motion about the third principal axes (note that, if a limiting motion of this type existed, it would be periodic with constant velocity). In this paper we derive feedback laws such that the closed-loop trajectories asymptotically approach the set

$$\Lambda = \{(\omega_1, \phi, \theta, \psi) : \omega_1 = 0, \phi = \theta = 0\}$$

(although a limiting motion may not necessarily exist or be periodic).

Theorem 4. Consider a rigid satellite with two actuators aligned along the first two principal axes. The feedback control law

$$\begin{aligned} u_1 = & -a_1 w_2 w_3 - w_1 - \phi - A_1 w_1 - B_1 w_1^2 \\ & - w_1 \cos \theta - w_1 \sin \theta - A_1 a_1 w_1 w_2 \\ & - 2B_1 a_1 w_1 w_2 w_3 \\ u_2 = & -a_2 w_1 w_3 - w_2 - \eta - A_2 w_1 - B_2 w_1^2 \\ & - w_1 \sin \theta \tan \phi - w_2 - w_1 \cos \theta \tan \phi \\ & - 2B_2 a_1 w_1 w_2 w_3 - A_2 w_1 w_2 \end{aligned}$$

where

$$A_1 A_2 = 0 \quad \text{and}$$

$$a_1(A_1(A_1 + B_2) - A_2(A_2 + B_1)) < 0$$

locally asymptotically stabilizes the rigid satellite model about a revolute motion about the third principal axis.

3. PROOF OF THEOREM 2

We shall use Krasnoselski's Theorem (Krasnoselski and Zabreiko, 1984) which gives necessary conditions for the origin to be locally asymptotically stable equilibrium for the autonomous differential equation

$$\dot{x} = F(x), \quad F(0) = 0, \quad x \in R^n \quad (3.1)$$

Explicitly, in order for (3.1) to be locally asymptotically stable, it is necessary that, locally,

$$\deg(-F) = 1$$

where \deg denotes the Brouwer degree. An immediate consequence, noted independently by Brockett (1983), is that the equation

$$F(x) = y, \quad \|y\| < \infty \quad (3.2)$$

must be solvable. In this section, we shall give an independent proof of Krasnoselski's Theorem and hence of the solvability of (3.2), but we will

first use this condition to prove Theorem 2.

Proof. Let $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in R^{n_1+n_2}$. For an arbitrary but fixed choice of static feedback laws $u_1(x_1, x_2), \dots, u_m(x_1, x_2)$ denote by $F(x)$ the closed-loop drift

$$\begin{aligned} F(x_1, x_2) = & \begin{pmatrix} F_1(x_1, x_2) \\ F_2(x_1, x_2) \end{pmatrix} \\ & f_1(x_1, x_2)x_1 + \sum_{i=1}^m b_i u_i(x_1, x_2) \\ & f_2(x_1, x_2) \end{aligned}$$

Now choose $\|y_1\| \ll \infty$ and $y_2 = 0$ and consider the solvability of equation (3.2). By (H2), we conclude that $x_1 = 0$ and therefore that $f_1(x_1, x_2)x_1 = 0$. Thus, solvability (3.2) reduces to solvability of

$$\sum_{i=1}^m b_i u_i(0, x_2) = y_1, \quad \|y_1\| \ll \infty$$

which can only occur if $m' = n_1$. In the case of dynamic feedback, the proof is identical.

We now turn to a proof of this necessary condition for local asymptotic stability.

Theorem 5. (Krasnoselski). A necessary condition for (3.1) to be locally asymptotically stable is

$$\deg(-F) = 1.$$

Proof. Choose a Lyapunov function $V(x)$ and a real number $c \ll \infty$. We remark that while it is tempting to believe that $V^{-1}(c)$ is diffeomorphic to a sphere, this is equivalent to the Poincaré conjecture [see Wilson (1967)]. Instead, we use another nontrivial assertion about the topology of Lyapunov functions:

Lemma. $V^{-1}(-\infty, c)$ is diffeomorphic to R^n . In particular, $M_c = V^{-1}(-\infty, c]$ is contractible.

Proof. The open subset $V^{-1}(-\infty, c)$ is smooth manifold with a vector field possessing a globally asymptotically stable equilibrium. According to Milnor (1964),

$$V^{-1}(-\infty, c) \approx R^n$$

so that $V^{-1}(-\infty, c]$ is contractible.

Returning to the proof of (3.1), denote by ϕ_T the time T -map induced by $F(x)$. According to the Lefschetz Fixed Point Formula (Griffiths and Harris, 1978), since $H^i(M_c) = \{0\}$ for $i \geq 1$,

$$\begin{aligned} L(\phi_T) &= \sum_{i=0}^n (-1)^i \operatorname{tr}(\phi_T^*: H^i(M_c) \rightarrow H^i(M_c)) \\ &= \operatorname{tr}(\phi_T^*: H^0(M_c) \rightarrow H^0(M_c)) = 1 \end{aligned}$$

where $L(\phi_T)$ is the oriented intersection number of $\text{graph}(\phi_T)$ with the diagonal of $M_1 \times M_1$. For $T \ll \infty$, such fixed points are the equilibria of $F(x)$, viz. 0. More explicitly, if F had a simple equilibrium, then (see Griffiths and Harris, 1978)

$$\deg(-F) = \deg_0(-F) = \text{ind}_0(F) = L(\phi_T). \quad (3.3)$$

Since degrees and intersection numbers are robust with respect to small perturbations, for a multiple equilibrium and for $T \ll \infty$, we conclude Krasnoselski's Theorem:

$$\deg(-F) = L(\phi_T) = 1. \quad (3.4)$$

4. GEOMETRIC CONTROL THEORETIC STABILIZATION TECHNIQUES

Although the full set of satellite equations with only two controls cannot be asymptotically stabilized via smooth feedback, one may still wish to design a control law that stabilizes at least a particular subset of them. The structure of the equations (2.4) is such that the behaviour of the state variables ω_1 , ω_2 , ω_3 and ϕ , θ is not directly affected by the value of the third Euler angle ψ (it could indeed be affected—indirectly—throughout a state feedback control). The state variables in question describe the motion of the satellite regardless of a rotation around one of the principal axis and one may wish to examine whether or not it is possible to achieve asymptotic stability of the motion at least up to a rotation around the axis.

Consider a rigid satellite with two actuators aligned along the first two principal axes. Suppose $a_3 = 1$ (this assumption will be relaxed later). Without loss of generality, we can suppose that, after feedback, the equations in question assume the form

$$\begin{aligned} \dot{\omega}_1 &= u_1 \\ \dot{\omega}_2 &= u_2 \\ \dot{\omega}_3 &= \omega_1 \omega_2 \\ \dot{\phi} &= \cos \theta \omega_1 + \sin \theta \omega_2 \\ \dot{\omega} &= \sin \theta \tan \phi \omega_1 + \omega_2 - \cos \theta \tan \phi \omega_3. \end{aligned} \quad (4.1)$$

The equations (4.1) can no longer be viewed as a system of the form (2.1) satisfying (H1)–(H3) and they are not feedback linearizable in a neighbourhood of 0. Thus, in order to obtain asymptotic stability, we have to use different techniques.

Consider again a nonlinear system having the form

$$\dot{x} = f(x) + \sum_{i=1}^m g_i(x) u_i, \quad x \in \mathbb{R}^n, \quad u_i \in \mathbb{R} \quad (4.2)$$

and suppose there exist smooth functions

$$y_i = h_i(x) \quad i = 1, \dots, m \quad (4.3)$$

defined in a neighbourhood of $x = 0$, satisfying the assumptions:

(H4) $h_i(0) = 0$ for $i = 1, \dots, m$ and

(H5) The matrix

$$A(x) = (a_{ij}(x)) = \left\{ \frac{\partial h_i}{\partial x_j} g_j(x) \right\} \quad (4.4)$$

is nonsingular at $x = 0$.

Because of (H5), the functions h_1, \dots, h_m have linearly independent differentials at $x = 0$ and it is possible to choose new local coordinates for (4.2) in the form

$$(z_1, \dots, z_{n-m}, y_1, \dots, y_m) = (z, y)$$

with $z_i(0) = 0$ for $i = 1, \dots, m$.

Rewritten in these coordinates, the system (4.2) becomes

$$\dot{z} = f_1(z, y) + g_1(z, y) u \quad (4.5a)$$

$$\dot{y} = f_2(z, y) + g_2(z, y) u \quad (4.5b)$$

with $f_1(0, 0) = 0$, $f_2(0, 0) = 0$ and $g_2(z, y)$ nonsingular at all (z, y) near $(0, 0)$.

The choice of feedback control law of the form

$$u = g_2^{-1}(z, y) [-y - f_2(z, y)] \quad (4.6)$$

changes (4.5) into a system of the form

$$\dot{z} = \tilde{f}_1(z, y) + \tilde{g}_1(z, y) y \quad (4.7a)$$

$$\dot{y} = -y \quad (4.7b)$$

where

$$\begin{aligned} \tilde{f}_1(z, y) &= f_1(z, y) - g_1(z, y) g_2^{-1}(z, y) [-y - f_2(z, y)] \\ \tilde{g}_1(z, y) &= -g_1(z, y) g_2^{-1}(z, y). \end{aligned}$$

It is apparent from (4.7) that locally around $x = 0$ the set

$$Z = \{x : h_i(x) = 0; i = 1, \dots, m\}$$

i.e. the set of points with coordinates $(z, 0)$ is an invariant manifold. The effect of the control law (4.6) is such as to constrain on Z every motion of (4.5) starting at a point of Z . On Z , the system evolves as

$$\dot{z} = \tilde{f}_1(z, 0). \quad (4.8)$$

In other words, (4.8) is a local description of the behaviour of the control system (4.2) under the constraint

$$h_i(x) = 0$$

A proper choice of the functions $h_1(x), \dots, h_m(x)$ can be helpful in solving a stabilization problem, as explained in the following statement.

Theorem 6. Consider a system (4.2) satisfying (H4)–(H5). Suppose, in addition, that

(H6) The constrained dynamics (4.8) has an asymptotically stable equilibrium at $z = 0$.

Then, the control law (4.6) locally stabilizes (4.5) at $(z, v) = (0, 0)$.

Proof. If the first variation of (4.8) has spectrum in the left-half complex plane, then the first variation of (4.7) being upper block-triangular also has its spectrum in the left half plane. If (4.8) is locally asymptotically stable but is critically stable, i.e. has some eigenvalues on the imaginary axis, then the constraint set Z contains a centre manifold for (4.7) [see Carr (1981)]. If Z is a centre manifold, then asymptotic stability of (4.7) follows immediately from the centre manifold theorem. If Z properly contains a centre manifold, then after a local change of coordinates, as in Byrnes and Isidori (1988), we can separate the state into two pieces one whose linearization is stable and which, when constrained to zero, defines the second, invariant, piece as a centre manifold. Again, by the centre manifold theorem, (4.7) is locally asymptotically stable.

Remark 1. Some early versions of stabilization schemes such as (4.6) were studied by Brockett (1983) under an additional "finite gain" hypothesis, which was later relaxed by Aeyels (1985) using centre manifold methods. Our approach uses output functions very heavily and in this way is inspired by classical frequency domain controls and its extensions to nonlinear, high gain controls via singular perturbations [see e.g. Young *et al.* (1977)] or sliding mode control [e.g. Utkin (1974)] with one important exception. Namely, for applications such as attitude stabilization of rigid spacecraft the motion on the sliding hypersurface is at best critically stable and cannot always be stabilized by high gain. For example, consider the system

$$\begin{aligned}\dot{z} &= y^2 - z^5 \\ \dot{y} &= z^2 + u.\end{aligned}$$

In Byrnes and Isidori (1989), it is shown that while the analogue of (4.6), *viz.*

$$u = -y - z^2$$

is locally stabilizing, for every value of k the high gain law

$$u = -ky$$

is destabilizing. In fact, no smooth output feedback law is locally stabilizing, for this choice of output. In the next section we show how to design appropriate outputs for the reduced rigid body model, analogous to control via "optimal" sensor location.

5. STABILIZATION OF THE SATELLITE EQUATIONS

We now return to the problem of stabilizing the satellite equations (4.1). The matter is to design "dummy" output functions

$$\begin{aligned}y_1 &= h_1(\omega_1, \omega_2, \omega_3, \phi, \theta) \\ y_2 &= h_2(\omega_1, \omega_2, \omega_3, \phi, \theta)\end{aligned}$$

in such a way that some of the procedures discussed in the previous section will be applicable.

Let us consider, for a moment, the simpler situation in which the last two equations of (4.1) are being neglected. Stabilizing control laws for these equations are already known, but we want to show how it is possible to derive them within the present approach.

In order to satisfy requirements (H4) and (H5), choose

$$\begin{aligned}h_1(\omega_1, \omega_2, \omega_3) &= -\omega_1 - f_1(\omega_3) \\ h_2(\omega_1, \omega_2, \omega_3) &= -\omega_2 - f_2(\omega_3)\end{aligned}$$

with $f_1(0) = f_2(0) = 0$.

The corresponding normal form (4.5) then becomes

$$\begin{aligned}\dot{y}_1 &= -\frac{\partial f_1}{\partial \omega_3}(f_1 f_2 + f_1 y_2 + f_2 y_1 + y_1 y_2) - u_1 \\ \dot{y}_2 &= -\frac{\partial f_2}{\partial \omega_3}(f_1 f_2 + f_1 y_2 + f_2 y_1 + y_1 y_2) - u_2 \\ \dot{\omega}_3 &= f_1 f_2 + f_1 y_2 + f_2 y_1 + y_1 y_2.\end{aligned}$$

The use of a feedback law of the form (4.6) yields a system

$$\begin{aligned}\dot{y}_1 &= -y_1 \\ \dot{y}_2 &= -y_2 \\ \dot{\omega}_3 &= f_1 f_2 + f_1 y_2 + f_2 y_1 + y_1 y_2\end{aligned}$$

whose stability depends entirely on that of

$$\dot{\omega}_3 = f_1(\omega_3)f_2(\omega_3).$$

From this, it is easily seen that the choice

$$\begin{aligned}f_1(\omega_3) &= \omega_3 \\ f_2(\omega_3) &= -\omega_3^2\end{aligned}$$

solves the problem. Note that the feedback law derived using these outputs and the method described in Section 4 coincides with the stabilizing control law found by Brockett (1983) [see also Aeyels (1985)].

The stabilization of the full set of equations (4.1) is not terribly more difficult. In order to fulfill requirements (F4) and (H5), set:

$$\begin{aligned}h_1(\omega_1, \omega_2, \omega_3, \phi, \theta) &= -\omega_1 - f_1(\omega_3, \phi, \theta) \\ h_2(\omega_1, \omega_2, \omega_3, \phi, \theta) &= -\omega_2 - f_2(\omega_3, \phi, \theta)\end{aligned}$$

with $f_1(0, 0, 0) = f_2(0, 0, 0) = 0$.

Using again a control law of the form (4.6), we are led to examine the asymptotic stability of the following equation:

$$\begin{aligned}\dot{\omega}_3 &= f_1(\omega_3, \phi, \theta) f_2(\omega_3, \phi, \theta) \\ \dot{\phi} &= -\cos \theta f_1(\omega_3, \phi, \theta) + \sin \theta \omega_3 \\ \dot{\theta} &= -\sin \theta \tan \phi f_1(\omega_3, \phi, \theta) - f_2(\omega_3, \phi, \theta) \\ &\quad - \cos \theta \tan \phi \omega_3.\end{aligned}\quad (5.1)$$

Since the linearization of (5.1) at 0 has the form (superscripts denote differentiation)

$$\begin{aligned}(\dot{\delta\omega}_3) &= 0 \\ (\dot{\delta\phi}) &= -f_1'''(0) \delta\omega_3 - f_1''(0) \delta\phi - f_1'(0) \delta\theta \\ (\dot{\delta\theta}) &= -f_2'''(0) \delta\omega_3 - f_2''(0) \delta\phi - f_2'(0) \delta\theta\end{aligned}$$

we are in the "critical case", the stability of which shall be analyzed via Centre Manifold Theory. In order to minimize the dimension of the centre manifold, we may wish to have the eigenvalues of

$$\begin{bmatrix} f_1''(0) & f_1'(0) \\ f_2''(0) & f_2'(0) \end{bmatrix}$$

in the left-half plane. This is accomplished, e.g. by means of the choice

$$\begin{aligned}f_1(\omega_3, \phi, \theta) &= \phi + g_1(\omega_3) \\ f_2(\omega_3, \phi, \theta) &= \theta + g_2(\omega_3).\end{aligned}\quad (5.2)$$

We now want to find g_1 and g_2 that make the dynamics (5.1) locally asymptotically stable at 0. To this end introduce a local change of variables and take the right-hand sides of (5.2) as new state variables, still denoted by f_1 and f_2 . The system (5.1) will thus be rewritten as

$$\begin{aligned}\dot{\omega}_3 &= f_1 f_2 \\ \dot{f}_1 &= -f_1 + \omega_3 \sin(f_2 - g_2) \\ &\quad + f_1(1 - \cos(f_2 - g_2)) + g_1''' f_1 f_2 \\ \dot{f}_2 &= -f_2 - f_1 \sin(f_2 - g_2) \tan(f_1 - g_1) \\ &\quad - \omega_3 \cos(f_2 - g_2) \tan(f_1 - g_1) + g_2''' f_1 f_2.\end{aligned}$$

This system has a local centre manifold $\{f_1 = z_1(\omega_3), f_2 = z_2(\omega_3)\}$ near zero (where $z_1(0) = z_2(0) = z_1'''(0) = z_2'''(0)$), which can be approximated by the solutions of

$$\begin{aligned}0 &= -z_1 + \omega_3(z_2 - g_2) + 0(\omega_3^4) \\ 0 &= -z_2 - \omega_3(z_1 - g_1) + 0(\omega_3^4).\end{aligned}$$

At this point, one has to make a choice of g_1, g_2 such that the solutions $z_1(\omega_3), z_2(\omega_3)$ of the previous equations render

$$\dot{\omega}_3 = z_1(\omega_3) z_2(\omega_3)$$

locally asymptotically stable. Setting e.g.

$$g_i = A_i \omega_3 + B_i \omega_3^2 \quad i = 1, 2$$

and taking Taylor series expansions for z_i

$$z_i = \alpha_i \omega_3 + \beta_i \omega_3^2 + 0(\omega_3^4) \quad i = 1, 2$$

one obtains

$$\begin{aligned}z_1 &= -A_2 \omega_3^2 + (A_1 - B_2) \omega_3^3 + 0(\omega_3^4) \\ z_2 &= A_1 \omega_3^2 + (A_2 + B_1) \omega_3^3 + 0(\omega_3^4).\end{aligned}$$

In order to have local asymptotic stability, provided that

$$A_1 A_2 \neq 0$$

it is sufficient to set

$$A_1(A_1 - B_2) = A_2(A_2 + B_1) = 0$$

Thus, a locally asymptotically stabilizing control law for (4.1) can be obtained when

$$\begin{aligned}h_1 &= -(\omega_3 + \phi + A + A_1 \omega_3 + B_2 \omega_3^2) \\ h_2 &= -(\omega_3 + \eta + A_2 \omega_3 + B_1 \omega_3^2)\end{aligned}$$

provided that $A_i, B_i, i = 1, 2$, satisfies the previous requirements. In case $a_i \neq 1$ the same arguments show that local asymptotic stability is achieved provided

$$a_i(A_1(A_1 - B_2) = A_2(A_2 + B_1)) = 0.$$

6. AN INSTABILITY RESULT

Consider again the model of a rigid satellite with just two controls and suppose its motion has been, as in the previous section, asymptotically stabilized up to a motion around one of the principle axes. The question that naturally arises is whether or not the state variable ψ_i , which measures the rotation around that axis, will eventually converge toward some finite limit as $t \rightarrow \infty$. If the Jacobian $[\partial u / \partial \phi, \partial u / \partial \eta]$ of a ψ -independent control law has rank two, then an argument similar to the one given in Section 5 shows that the closed-loop system has the one-dimensional manifold

$$\Lambda = \{(\omega_i, \phi, \theta, \psi_i) : \omega_i = 0, \phi = \theta = 0\} \quad (6.1)$$

as its equilibrium set and it makes sense to ask (locally) whether closed-loop trajectories tend, in fact, to a point of Λ .

More generally, throughout the rest of this section we shall assume that all feedback laws we consider satisfy:

(H7) The equilibria of the closed-loop system are given by (6.1)

The questions we shall study is whether feedback laws satisfying (H7) exist which robustly stabilize the closed loop system.

Definition 1. A feedback law $u_i(\omega, \phi, \theta, \psi)$ satisfying (H7) robustly stabilizes (4.1) provided for all ϵ sufficiently small, the feedback law

$$u_i(\omega, \phi, \theta, \psi) + \epsilon v_i(\omega, \phi, \theta, \psi)$$

renders the closed loop system (Lyapunov) stable.

In order to analyze this question we first note that from the form of the closed-loop equations, a robustly stabilizing feedback law must have no eigenvalues of the (partial) Jacobian (evaluated at 0)

$$\begin{array}{cc} -\frac{\partial u_1}{\partial \phi} & \frac{\partial u_1}{\partial \theta} \\ \frac{\partial u_2}{\partial \phi} & \frac{\partial u_2}{\partial \theta} \end{array} \quad (6.2)$$

being zero. Henceforth, we may assume that (6.2) is nonsingular.

Without loss of generality, consider the equilibrium point $\omega_1 = \omega_2 = \omega_3 = \phi = \eta = \psi = 0$. The linearized system around 0 has two eigenvalues at the origin and four eigenvalues with negative real parts.

Thus, the system has a two-dimensional centre manifold at 0. Set

$$z = (z_1, z_2, z_3, z_4) = (\omega_1, \omega_2, \phi, \theta)$$

and recall that

$$u_1(0, 0, 0, 0, 0, \psi) = u_2(0, 0, 0, 0, 0, \psi) = 0$$

for all ψ near 0. The (controlled) satellite dynamics has the form

$$\begin{aligned} \dot{z} &= Az + B\omega_3 + h(z, \omega_3, \psi) \\ \dot{\omega}_3 &= z_1 z_3 \\ \dot{\psi} &= \omega_3 + z_1 g_1(z_3, z_4) + \omega_3 g_1(z_3, z_4) \end{aligned} \quad (6.3)$$

where A is a matrix with all the eigenvalues in the left-half plane, $h(z, \omega_3, \psi)$ is a function that does not contain linear terms and is such that

$$h(0, 0, \psi) = 0$$

for all ψ near zero, $g_1(0, 0) = g_3(0, 0) = 0$.

A linear change of variables

$$\xi = Tz + k\omega_3$$

where T is an invertible matrix, brings (6.3) to the form

$$\begin{aligned} \dot{\xi} &= \hat{A}\xi + \hat{h}(\xi, \omega_3, \psi) \\ \dot{\omega}_3 &= (c_1\xi + D_1\omega_3)(c_2\xi + D_2\omega_3) \\ \dot{\psi} &= \omega_3 + \hat{g}(\xi, \omega_3) \end{aligned}$$

where $\hat{h}(0, 0, \psi) = 0$ for all ψ near zero and $\hat{g}(\xi, \omega_3)$ does not contain linear terms. A centre manifold for this system at 0 is the graph of a

function $\xi = \pi(\omega_3, \psi)$ such that $\pi(0, 0) = \pi^{\omega_3}(0, 0) = \pi^\psi(0, 0) = 0$. Such a function is a solution of the partial differential equation

$$\begin{aligned} \pi^{\omega_3}(c_1\pi + D_1\omega_3)(c_2\pi + D_2\omega_3) \\ + \pi^\psi(\omega_3 + \hat{g}(\pi, \omega_3)) = \hat{A}\pi + \hat{h}(\pi, \omega_3, \psi). \end{aligned}$$

Since $\hat{h}(0, 0, \psi) = 0$ for all ψ near zero, we have necessarily for any centre manifold

$$\pi(0, \psi) = 0$$

for all ψ , and this in turn implies

$$\pi(\omega_3, \psi) = \omega_3 \hat{\pi}(\omega_3, \psi).$$

As a consequence, the stability of the flow on the centre manifold near zero is governed by equations [see (6.3)] of the form

$$\begin{aligned} \dot{y} &= y^2 P(x, y) \\ \dot{x} &= y + yQ(x, y) \end{aligned} \quad (6.4)$$

where $Q(0, 0) = 0$.

An appropriate choice of Lyapunov function shows that $x = y = 0$ is an unstable equilibrium of (6.4). Set

$$V(x, y) = -xy$$

and note that

$$\dot{V}(x, y) = -y^2(1 + Q(x, y) + xP(x, y)).$$

If $r > 0$ is sufficiently small, on the open disc

$$D_r = \{(x, y) : x^2 + y^2 < r^2\},$$

we have

$$|Q(x, y) + xP(x, y)| < 1.$$

At each point of D_r where $V(x, y) < 0$, we also have $\dot{V}(x, y) < 0$ and therefore the equilibrium is unstable.

We can thus conclude the following.

Theorem 7. There is no smooth state feedback law, having closed-loop equilibrium (6.1), robustly stabilizing a rigid satellite using two gas-jet actuators.

In closing, we note [see e.g. Xu (1986)] that the open-loop system (2.4) is actually unstable, showing that the instability result, Theorem 7, holds in a broader context. It is tempting to conjecture that no smooth feedback law exists, stabilizing (in the sense of Lyapunov) a rigid satellite with two controls.

7. CONCLUSIONS

Important questions such as controllability, reachability and feedback stabilizability for the standard rigid body model of a spacecraft controlled by three independent pairs of gas-jets

can be analyzed elegantly and simply using the recent method of nonlinear feedback linearization. For the case of a spacecraft with two independent controls, modelling for example failure of an actuator, local controllability and reachability have been known to hold for almost all actuator configurations but the question of feedback stabilization has heretofore remained a heavily researched yet unanswered question. Using topological methods, it is shown that such systems cannot be stabilized by smooth feedback. This is also a special case of a more general result asserting that for certain classes of systems feedback stabilization can be achieved precisely when feedback linearization is possible. On the other hand, using some general feedback design methods which comprise a nonlinear enhancement of root-locus techniques, it is possible to derive explicit feedback laws stabilizing this system about an attractor, inducing a closed-loop system with trajectories tending to a revolute motion about a principal axis. The general problem of feedback stabilization about attractors appears to be an important extension of stabilization about an equilibrium, yielding bounded trajectories when stabilization about an equilibrium is not possible.

Acknowledgements—C. L. Byrnes's research is partially supported by AFOSR, NSF, and A. Isidori's research is partially supported by Telespazio, S.P.A. and the Ministero della Pubblica Istruzione.

REFERENCES

- Aevels, D. (1985) Stabilization of a class of nonlinear systems by a smooth feedback. *Syst. Control Lett.* **5**, 289–294.
- Brockett, R. W. (1983) Asymptotic stability and feedback stabilization. In R. W. Brockett, R. S. Millmann and H. Sussmann (Eds), *Differential Geometric Control Theory*, pp. 181–191. Birkhäuser, Boston.
- Brockett, R. W. (1978) Feedback invariants for nonlinear systems. *6th IFAC Congress*, pp. 1115–1120.
- Byrnes, C. and A. Isidori (1984) A frequency domain philosophy for nonlinear systems, with application to stabilization and to adaptive control. *23rd IEEE Conf. on Decision and Control*, pp. 1569–1573.
- Byrnes, C. and A. Isidori (1986) Asymptotic expansions, root-loci and the global stability of nonlinear feedback systems. In M. Fliess and M. Hazewinkel, (Eds), *Algebraic and Geometric Methods in Nonlinear Control Theory*, pp. 159–179. Reidel, Hingham, MA.
- Byrnes, C. and A. Isidori (1985) Global stabilization of nonlinear minimum phase systems. *24th IEEE Conf. on Decision and Control*.
- Byrnes, C. and A. Isidori (1988a) Heuristics for nonlinear control. In C. Byrnes and A. Kurzhansky (Eds), *Modelling and Adaptive Control*, Septon 1988. LNCS **185**, pp. 48–70. Springer, Berlin.
- Byrnes, C. and A. Isidori (1988b) Local stabilization of minimum phase systems. *Syst. Control Lett.* **11**, 9–17.
- Byrnes, C. and A. Isidori (1989) New results and examples in nonlinear feedback stabilization. *Syst. Control Lett.* **12**, 437–442.
- Carr, J. (1981) *Applications of Center Manifold Theory*. Springer, New York.
- Crouch, P. E. (1984) Spacecraft attitude control and stabilization: Applications of geometric control to rigid body models. *IEEE Trans. Aut. Control*, **AC-29**, 321–331.
- Crouch, P. E. (1985) Attitude control of spacecraft. *Mathematical Control Theory. Banach Center Publications*, **14**, 121–134.
- Crouch, P. E. and M. Trivelpiece (1983) On sufficient conditions for local asymptotic stability of nonlinear systems whose linearization is uncontrollable. *Control Theory Centre Report No. 114*, Univ. of Warwick, U.K.
- Griffiths, P. and J. Harris (1978) *Principles of Algebraic Geometry*. Wiley, New York.
- Hermes, H. (1980) On a stabilizing feedback attitude control. *SIAM J. Control Optim.* **18**, 352–361.
- Ha, X. (1986) Stability of nonlinear feedback systems in the critical case. M. Sc. Thesis, ASU.
- Hunt, L. R., R. Su and G. Meyer (1983) Design for multi-input nonlinear systems. In R. W. Brockett, R. S. Millmann and H. Sussmann (Eds), *Differential Geometric Control Theory*, pp. 268–298. Birkhäuser, Boston.
- Jakubczyk, B. and W. Respondek (1980) On linearization of control systems. *Bull. Acad. Polonaise Sci. Ser. Sci. Math.* **28**, 517–522.
- Krasnoselski, M. A. and P. P. Zabreiko (1984) *Geometrical Methods of Nonlinear Analysis*. Springer, Berlin.
- Marino, R. (1985) High gain feedback in nonlinear control systems. *Int. J. Control*, **42**, 1369–1385.
- Milnor, J. W. (1963) Differential topology. I. J. Saaty (Ed.), *Lectures in Modern Mathematics*. Wiley, New York.
- Sommer, R. (1980) Control design for multivariable nonlinear time-varying systems. *Int. J. Control*, **31**, 883–891.
- Sontag, E. D. and H. J. Sussmann (1980) Remarks on continuous feedback. *Proc. 19th IEEE Conf. on Decision and Control*, Albuquerque.
- Utkin, V. I. (1974) *Sliding Modes and their Use in Variable Structure Systems*. Nauka, Moscow.
- Wilson, F. W. Jr (1967) The structure of the level surfaces of a Lyapunov function. *J. Diff. Equ.* **4**, 323–329.
- Young, K. K. D., P. V. Kokotovich and V. I. Utkin (1977) A singular perturbation problem of high gain feedback systems. *IEEE Trans. Aut. Control*, **AC-22**, 931–937.
- Zabrezyk, J. (1989) Some comments on stabilizability. *Applied Math. Optim.* **19**, 1–9.

Knowledge Engineering for Industrial Expert Systems*

GUNNAR JOHANNSEN† and JAMES L. ALTY‡

Knowledge engineering is split into knowledge acquisition and system implementation, and a review provides examples of knowledge elicitation and machine induction in industrial domains.

Key Words: Knowledge engineering, knowledge acquisition, knowledge elicitation, machine induction, expert systems, supervision and control, industrial control, cognitive task analysis, qualitative modelling, knowledge acquisition tools.

Abstract—The inherent difficulties involved in the process of extracting knowledge from experts are discussed and identified. Such difficulties have resulted in few expert systems progressing beyond the prototyping stage. The conflicting terminology used to describe the whole process is examined and, as a result, knowledge engineering is defined as the appropriate term for the whole process. This is then further split into knowledge acquisition and system implementation. Finally, knowledge acquisition is further subdivided into knowledge elicitation and machine induction.

The particular problems associated with the construction of expert systems in industrial control applications are discussed. Such systems are characterised by the nature of their user population, the type of support provided and whether they operate on-line or off-line. The importance of defining functionality and goals at the outset is stressed. The need for user models is also highlighted. The various techniques used in knowledge elicitation—interviews, questionnaires, observations, protocol analyses, teachback, interviewing, walkthroughs and formal techniques—are briefly reviewed. The alternative approach using machine induction techniques is also discussed. An examination is made of the competing approaches involving bottom-up and top-down techniques. The benefits resulting from the application of cognitive task analyses rather than technology-driven approaches are also stressed. Current knowledge acquisition tools such as KRITON, KADS, ACQUISIT, KEAIS and ROGET are reviewed.

Examples are given of the use of time-line techniques in power plant knowledge acquisition, knowledge and task analyses in the construction of a failure management expert system and of the use of inductive techniques in gas oil separator design and satellite power systems control. In the latter case, the use of qualitative modelling is highlighted.

The possibility of domain experts in industrial control carrying out their own knowledge engineering is examined but rejected as unlikely, unless better tools exist. The provision of better tools is identified as one of the key factors required to simplify the knowledge engineering process.

INTRODUCTION

KNOWLEDGE ENGINEERING is the process of building expert systems. Such systems are medium- to large-scale software products which are designed to solve problems of different kinds using a knowledge-based approach where the knowledge is represented in an explicit manner. They have a wide area of applicability particularly in industrial control. Hayes-Roth *et al.* (1983), for example, have identified 10 generic categories of knowledge engineering applications. These are interpretation, prediction, diagnosis, design, planning, monitoring, debugging, repair, instruction, and control. Such systems normally contain two main components (Davies, 1982): the inference mechanism (the problem solving component) and the knowledge base (which may actually comprise a number of knowledge bases). Generally speaking, expert systems work best in narrow application domains.

Madni (1988) has provided a hierarchical classification of expert systems from a human factors perspective. One level of his classification distinguishes between expert systems with respect to different purposes:

- perform a task;
- assist in a task; and
- teach a task.

The first category deals with autonomous expert systems such as those found in autonomous robots or automation systems. The second and third categories are concerned with expert

* Received 30 July 1989; revised 11 February 1990; received in final form 17 February 1990. The original version of this paper was presented at the 4th IFAC/IFIP/IFORS/IEA Conference on Man-Machine Systems: Analysis, Design and Evaluation which was held in Xi'an, People's Republic of China during September, 1989. The published Proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Editor A. P. Sage.

† Laboratory for Man-Machine Systems, University of Kassel (GhK), D-3500 Kassel, Federal Republic of Germany. Author to whom all correspondence should be addressed.

‡ Scottish Human-Computer Interaction Centre, University of Strathclyde, Glasgow, Scotland, U.K.

consultation systems. In addition, systems concerned with teaching have more granular uncompiled knowledge and access a pedagogical knowledge base in addition to domain knowledge bases. In industrial systems, expert systems of the first category are always embedded within the technical system whereas those in the second category can be either stand-alone or embedded.

The process of building an expert system consists of two main activities which usually overlap—acquiring the knowledge and implementing the system. The acquisition activity involves the collection of knowledge about facts and reasoning strategies from the domain experts. Usually, such knowledge is elicited from the experts by so-called knowledge engineers, using interviewing techniques or observational protocols. However, machine induction, which automatically generates more elaborate knowledge from an initial set of basic knowledge (usually in the form of examples), has also been extensively used (Michie and Johnston, 1985). In the system construction process, the system builders (i.e. knowledge engineers), the domain experts and the users work together during all stages of the process, which traditionally has involved extensive prototyping.

To automate the problem solving process, the relevant task knowledge in the domain of interest needs to be understood in great detail but acquiring the knowledge for expert system building is generally regarded as a hard problem. This is not surprising. As Kidd (1987) has pointed out, acquiring knowledge from an expert entails answering some really fundamental questions such as

- what is the relationship between knowledge and language?
- how can we characterise different domains?
- what constitutes a theory of problem solving?

Clancey (1986) has also pointed out that the process of extracting knowledge from an expert is not the process of transferring a mental model lying in the brain of an expert into the mind of the system builder, but the formalisation of a domain for the first time, and this is inherently a difficult process. Ideally, models of conceptual structures of problem solving behaviour are required as a prerequisite to the knowledge transfer process. However, cognitive science approaches have not yet yielded sufficient information to enable a full understanding of the knowledge structures and problem solving strategies of experts to be applied, so that current approaches are incomplete and often *ad-hoc*. Rouse *et al.* (1989) view the situation of being able to capture human skills and

knowledge as possibly improving but conclude also that there may even be limits to our understanding when searching for models of conceptual structures.

The situation is further complicated by the fact that experts often have faulty memories or provide inconsistencies. This means that separate validation of the expertise elicited from experts is essential (Chignell and Peterson, 1988). Furthermore, experts exhibit cognitive biases such as overconfidence, simplification, and a low preference for the abstract, the relative and conflicting evidence. It is therefore important to test and validate expert systems both by analysing the expertise in the knowledge base and by examining failures in actual performance. As far as possible, cognitive biases should be filtered out during the elicitation process.

Much experimental evidence exists about the limitations of human decision making and it has been suggested that the development of systems which mimic human problem solving should be approached with some degree of caution (Tolcott *et al.*, 1989). In order to reduce the chances of bias, experts should be made aware of commonly found biases in judgement, the elicitation process should include probes to foster the consideration of alternatives and when experts run through sample problems in the elicitation process it should be borne in mind that the way in which the problems are presented will have an impact as to how far any derived rules will exhibit cognitive bias.

Madni (1988) has taken these important points into account in his detailed view of the whole knowledge engineering process appraised from a cognitive engineering viewpoint. He has suggested the following six stages which he terms mainstream development:

1. knowledge elicitation
2. cognitive bias filtering
3. knowledge representation and control scheme selection
4. software development and integration
5. system evaluation and validation
6. advanced prototype expert system.

Stages three and four ideally should only be carried out after the elicitation and cognitive bias stages have been completed. In reality, this is not possible and our own experience as well as that of other researchers suggests that several iterations through the first five stages are required before stage six can be contemplated. Madni also proposes two additional paths of prototyping activities for demonstration and software development purposes which are to be

performed in parallel to the first four stages of the mainstream development. Evaluations have to be carried out in all stages of software development.

It is important to realise that experts change their solution strategies dependent upon the boundedness of the problem (Mullin, 1989). In well-bounded problems an expert's approach differs dramatically from that of a novice. Experts' conceptual models reflect the physical processes that actually occur. By contrast the models used by nonexperts do not account for all the process parameters driving the problem. When problems concerning the processes are less well understood, expert and novice models appear superficially to be similar, though the experts seem to recognise that the simple models are not accurate and that the use of a precise model is not practicable. So experts know what they do not know and can readily identify features of uncertainty that preclude the use of precise solution strategies and in such circumstances will adopt simple and somewhat inaccurate process models. Knowledge acquisition techniques must be able to cope with this variation in expert strategy.

Few systems have progressed beyond the research or prototype phase mainly because of the inherent difficulties in the knowledge acquisition process (Breuker and Wiehnga, 1987). However, there is strong evidence that expert systems are now becoming cost effective. A recent issue of *Expert Systems Strategies* (Harmon, 1989) identified 30 successful systems selected from the contents of the First Conference on Innovative Applications of Artificial Intelligence sponsored by AAAI. The applications included a Packaging Advisor from Du Pont, a Labor Management Package from Ford, a Design Advisor from NCR and a Planning and Scheduling System for the US Navy. Comments on savings for various systems (not necessarily the above) were "savings of over \$1 million", "a 30% saving", "\$250K saved", "\$2 million per year in direct payback", "50% reduction in planning time", "a ten-fold speed up", "50% reduction in design time", and "tens of thousands of dollars saved". It appears that a more realistic approach to what is possible in expert systems technology coupled with better development tools is now beginning to yield the long awaited pay-back.

THE TERMINOLOGY OF KNOWLEDGE ENGINEERING

There are a number of terms used to describe the expert system building process which are not well defined and appear to overlap. Such terms

include knowledge elicitation, knowledge acquisition, system implementation, machine induction and even the term knowledge engineering itself. Buchanan *et al.* (1983) define knowledge acquisition as "the transfer and transformation of problem-solving expertise from some knowledge source to a program". This definition covers the whole process including identification of the problem, its conceptualisation, formalisation, implementation, testing and prototype revision. Diederich and Linster (1989) subdivide knowledge acquisition into knowledge elicitation and an operational phase. Motta *et al.* (1989) term the whole process knowledge engineering but subdivide it into knowledge acquisition, knowledge representation and implementation. They further break down knowledge acquisition into knowledge elicitation and data interpretation. As Motta *et al.* state, "The separation of acquisition from implementation leads to a view of knowledge acquisition as the production of an abstract architecture distinct from the implementation of the system". However, they accept that such a characterisation is also problematic since the only way of testing the knowledge is to run it so that the boundaries between acquisition and implementation can be very fuzzy.

Our view is that the process of building knowledge-based systems is essentially one of knowledge engineering and we regard the different terms as fitting together as in Fig. 1.

Most authors agree over the general term knowledge engineering. We have, however, distinguished the knowledge acquisition process from the system implementation process [like Motta *et al.* (1989)]. Although it is true that the two intertwine during prototyping, there are good reasons for separating them out, at least conceptually, as will become clear later on. We have also separated out elicitation (either manual or automatic) from machine induction since these acquisition techniques are quite distinct and have followed different development paths.

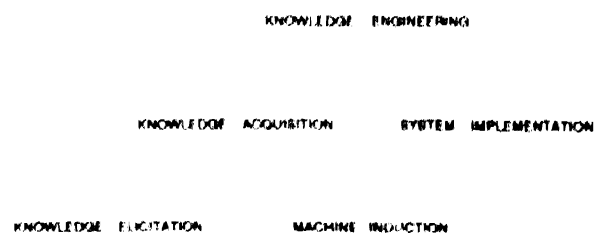


Fig. 1. Relationship between terms in knowledge engineering

KNOWLEDGE ENGINEERING IN INDUSTRIAL SYSTEMS

Expert systems can be introduced into industrial systems to provide support for different classes of people such as designers, operators and maintenance personnel. In general such systems will be off-line (for designers and maintenance personnel) and on-line (for operators). The knowledge engineering task will be different for each of these applications since the tasks involved will comprise different knowledge sources and structures. One difference is that between technological/scientific knowledge and experimental knowledge. This difference was described as a "knowledge of functioning" versus a "knowledge of utilisation" by DeMontmollin and DeKeyser (1986). The former knowledge is used by designers and maintenance personnel whereas the latter characterises that used by operators. The adoption of a truly human-centred design approach (Alty and Johannsen, 1989; Johannsen, 1990) requires designers to consider both. This means that user models are expected to be part of future knowledge-based support functions for designers (Sundström, 1988; Sundström and Johannsen, 1989).

Generally, designer activity in industrial systems ranges from the computer-aided design of subsystems and components to picture design for visual display units in control rooms (Rouse 1986; Elzer and Johannsen, 1988). Thus for designer support, knowledge about the application domain in addition to that concerned with design procedures is needed for all these tasks (Borys, 1988). Further, ergonomic knowledge and knowledge about user behaviour is required in picture design. Hence, design activity should be supported with a set of different knowledge bases (and perhaps a user model).

Off-line knowledge-based systems are not time critical. They may utilise several knowledge sources including technical documents, reference literature, handbooks, ergonomic knowledge, and knowledge about operator personnel (for use in user modelling). Whilst their operation is not time critical they may still have to take into account operator time constraints.

The most critical and challenging industrial expert systems are those developed for system operation. They may encompass support for the automatic system as well as support for the operators and may provide heuristic control, fault diagnosis, consequence prediction, and procedural support (Johannsen, 1990). The latter is particularly suitable for consistency checking of input sequences or for operator intent recognition (Hollnagel, 1987; Shalin *et al.*,

1988). All these support expert systems work under time constraints because they are running in parallel with the dynamic industrial process. Like off-line systems, these expert systems will depend upon a number of knowledge sources related to knowledge of functioning and knowledge of utilisation. Additional knowledge such as that of senior engineers will be required.

Whilst a support expert system for predicting the consequences of some technical failure will normally need only engineering knowledge, procedural support, diagnosis and heuristic control modules will need operational knowledge as well. Since they will also have to be integrated with the supervision and control system they will need to support numerical as well as symbolic knowledge.

The importance of signal processing as well as symbol processing has been emphasised by Rouse *et al.* (1989). They point out that models of symbol processing are much harder to be identifiable than those of signal processing because semantics and pragmatics play a large role in symbol processing systems. The need for symbolic representations is particularly underlined for industrial process control applications.

In all cases of knowledge-based systems developments, it will be necessary to define carefully the goals and functionalities of the various systems and their interdependencies at an early stage. It is also important to realise that in the industrial environment not all applications are suitable for the application of knowledge-based techniques. For example, existing numerical supervision and control systems are based upon thorough engineering methodologies and replacement by knowledge-based techniques would in most cases lead to performance degradation.

Finally, it must be realised that most industrial applications are very complex and this makes the problems of acquiring and assembling the knowledge in the industrial environment much more severe than in traditional computing domains. The elicitation and conceptualisation processes are liable to be far more complex and attempts to prove the consistency of the knowledge will be very time-consuming (Shirley, 1987; Johannsen, 1989). The full process is likely to take years rather than months. In the absence of a powerful methodology, we will have to work with inadequate tools for some time to come.

TECHNIQUES FOR KNOWLEDGE ACQUISITION

The techniques used in knowledge acquisition can be broadly divided into two categories—elicitation and machine induction. Strictly

speaking, there is a continuum between human-human elicitation and automatic induction. Three general principles have been proposed for the acquisition process by Gruber and Cohen (1987). They are concerned with primitives and generalisations.

The first principle prescribes that task-level primitives should be designed in order to capture important domain concepts defined by the expert. The knowledge engineer must use a language of task-level terms rather than imposing implementation-level primitives. This principle stresses the importance of separating out acquisition from implementation and will be referred to again later in the paper. These task-level primitives must be natural constructs for describing information, hypotheses, relations, and actions, in the domain expert's language. This would suggest that task analyses should be combined with knowledge analyses (Borys *et al.*, 1987; Johannsen, 1989).

The second principle suggests that explicit declarative representational primitives are preferable to procedural descriptions. This principle is based upon the observation that most experts more easily understand declarative representations. Formulating procedural aspects in this way "can facilitate acquisition, explanation, and maintenance" (Gruber and Cohen (1987) suggest that an expert should be asked "for the parameters of a domain that affect control decisions, and then to formulate control knowledge in terms of these parameters").

The third principle requires representations at the same level of generalisation as the expert's knowledge. Experts should not be forced to generalise except when absolutely necessary and they should not be asked to specify information not available to them. An example of an oversimplified generalisation would be the requirement to categorise a process variable as high, medium or low, when the expert needs to differentiate between many more steps or even a full range of numbers.

Knowledge elicitation

A number of techniques for knowledge elicitation are now in use. They usually involve the collection of information from the domain expert(s) either explicitly or implicitly. Originally, reports written by the experts were used but this technique is now out of favour since such reports tend to have a high degree of bias and reflective thought. Current techniques include interviews (both structured and unstructured), questionnaires or observational techniques such as protocol analyses and walkthroughs.

As pointed out by Forsythe and Buchanan (1989), knowledge elicitation methodologies "have more in common with the field-work orientation of anthropology and qualitative sociology than with the experimental orientation of many in the cognitive sciences". It is suggested that knowledge engineers also use the large amount of literature and experience as well as the much longer tradition of the social sciences in field work, particularly data-gathering methods such as face-to-face interviewing. Some pitfalls of knowledge elicitation are described on the basis of this experience in the social sciences. In particular, some interviewing problems such as obtaining data versus relating to the expert as a person, fear of silence and failing to listen, difficulty in asking questions, and interviewing without a record as well as conceptual problems such as treating interview methodology as unproblematic or blaming the expert are explained.

Interviews. In a structured interview, the knowledge engineer is in control. Such interviews are useful for obtaining an overall sense of the domain. In an unstructured interview, the domain expert is usually in control; however, such interviews can, as the name implies, yield a somewhat incoherent collection of domain knowledge. The result can be a very unstructured set of raw data that needs to be analysed and conceptualised. It is obviously important for the knowledge engineer to have some knowledge of the domain before wasting the valuable time of the expert. This might be obtained through textbooks, manuals and other well-documented sources. Group interviews can be useful particularly in the phase of cognitive bias filtering.

Questionnaires and rating scales. Questionnaires can be used instead or in addition with interviews. The interviews can be standardised in question-answer categories or questionnaires can be applied in a more formal way. However, the latter should be handled in most cases in a relaxed manner for reasons of building up an atmosphere of confidence and not disturbing the expert too much when applied in actual work situations (Borys *et al.*, 1987).

Rating scales are formal techniques for evaluating single items of interest by asking the expert to cross-mark a scale. Verbal descriptions along the scale such as from "very low" to "very high" or from "very simple" to "very difficult" are used as a reference for the expert. The construction, use and evaluation of rating scales is described very well in the psychological and social science literature. Rating scales can also be combined with interviews or questionnaires.

Observations. Observations are another technique for knowledge elicitation. They require little or no active participation of the expert. All actions and activities of the expert are observed as accurately as possible by the knowledge engineer who makes recordings of all the observed information. A special mixture of interview and observation techniques are the observation interviews (Matern, 1984; Johannsen, 1989). Sequences of activities are observed and questions about causes, reasons and consequences asked by the knowledge engineer during these observations. The combined technique is very powerful because the sequence of activities is observable whereas decision criteria, rules, plans etc. are elicited in addition through what-, how- and why-questions.

Protocol analysis. Protocol analyses are useful for obtaining detailed knowledge. It can involve verbal protocols in which the expert thinks aloud whilst carrying out the task, or motor protocols in which the physical performance of the expert is observed and recorded (often on videotape). Eye movement analysis is an example of a very specialised version of this technique. Motor protocols, however, are usually only useful when used in conjunction with verbal protocols.

In a verbal protocol, the expert thinks aloud and a time-stamped recording is made of his utterances (Ericsson and Simon, 1984). In such protocols, the expert should not be allowed to include retrospective utterances. He or she should avoid theorising their behaviour and should "only report information and intentions within the current sphere of conscious awareness" (Newell and Simon, 1972). As a verbal protocol is transcribed, it is broken down into short lines corresponding roughly to meaningful phrases [see Kuipers and Kassirer (1987) for examples of the technique]. The technique can collect the basic objects and relations in the domain and establish causal relationships. From these a domain model can be built. The experience with the use of verbal protocols from the analysis of trouble-shooting in maintenance work of technicians is described by Rasmussen (1984).

It is important when using the transcription method not to allow any proposed expert systems technology (i.e. rule-based approach) to influence the selection of items. Fox, when examining failures in the performance of an expert system designed to diagnose leukemia, noted that the expert systems technology used (in this case EMYCIN) strongly influenced the method used to "identify" useful information in the verbal protocols (Fox *et al.*, 1987). He also comments "We are even less confident about

knowledge that may be implicit or distributed in the structure of the protocols rather than concentrated in identifiable fragments".

The critical decision method (CDM) as described by Klein *et al.* (1989) is a special protocol analysis which elicits knowledge from experts and novices in a retrospective way. Nonroutine cases such as critical incidents are selected in order to discriminate the true expert's knowledge. Sources of bias are minimised by asking for uninterrupted incident descriptions. Then the history of the incident is reconstructed by means of time lines and decision points are identified and probed. It is stated that knowledge can be elicited with relatively little effort by using the critical decision method.

Teachback interviewing. In this technique, the expert first describes a procedure to the knowledge engineer, who then teaches it back to the expert in the expert's terms until the expert is completely satisfied with the explanation. Johnson and Johnson (1987) describe this technique and illustrate its use in two case studies. Their approach is guided by Conversation Theory (Pask, 1974), in which interaction takes place at two levels—specific and general. The paper gives a useful set of guidelines on the strengths and weaknesses of the technique.

Walkthroughs. More detailed than protocol analysis, and often better because they can be done in the actual environment which gives better memory cues. They need not, however, be carried out in real time. Indeed, such techniques are useful in a simulated environment where states of the system can be frozen and additional questions pursued.

Time lines. Tables in which several items of knowledge are contained in columns. The left column has to be filled with the time of occurrence of particularly interesting events such as failures or operator actions. Related information about the behaviour of the technical process, the automatic system and the human operators, at these times is recorded in separate columns with as much detail as is felt appropriate (Johannsen, 1989).

Formal techniques. These include multidimensional scaling, Repertory Grids and hierarchical clustering. Such techniques tend to elicit declarative knowledge. The most commonly used is the Repertory Grid Technique (Kelly, 1955) based on personal construct theory. It is used in ETS (Boose, 1986) which assists in the elicitation of knowledge for classification type problems, and PLANET (Shaw and Gaines, 1986). In ETS, the expert is interviewed to obtain elements of the domain. Relationships are then established by presenting triads of

elements and asking the expert to identify two traits which distinguish the elements. These are called constructs. They are then classified into larger groups called constellations. Various techniques such as statistical, clustering and multidimensional scaling are then used to establish classification rules which generate conclusion rules and intermediate rules together with certainty factors. The experts are interviewed again to refine the knowledge. ETS is said to save 2-5 months over conventional interviewing techniques. The system has been modified and improved and is now called AQUINAS (Boose and Bradshaw, 1988). To obtain procedural knowledge, techniques such as verbal protocols should be used.

Machine induction

Machine induction is a special case of machine learning which encompasses heuristics for generalising data types, candidate elimination algorithms, methods for generating decision trees and rule sets, function induction and procedure synthesis. A framework has been developed for describing such techniques that allows an evaluation of the usefulness of any technique to particular knowledge engineering problems (MacDonald and Witten, 1989). We have concentrated upon decision tree and rule set generation approaches because these techniques have been successfully used in a number of knowledge acquisition situations.

It is a common observation that experts have great difficulty in explaining the procedures which they use to arrive at decisions. Indeed, experts often make use of assumptions and beliefs which they do not explicitly state, and are surprised when the consequences of these hidden assumptions are pointed out (Jackson, 1985). The inductive approach relies on the fact that experts can usually supply examples of their expertise even if they do not understand their own reasoning mechanisms. This is because creating an example set does not require any understanding of how different evidence is assessed or what conflicts were resolved to reach a decision. Sets of such examples are then analysed by an inductive algorithm [one of the most popular being the ID3 algorithm of Quinlan (1979)] and rules are generated automatically from these examples.

The problem with inductive techniques is that the rules induced depend both upon the example set chosen and the inductive algorithm used. There is no guarantee that the rules induced will be valid knowledge. The approach therefore normally involves a checking with the expert to see if the induced rules are reasonable. It is not

uncommon to cycle a number of times through the induction process refining the knowledge base with the domain expert. Bratko (1989) gives a useful account of the techniques and the application of the ID3 algorithm. Hart (1987) has given guidelines on the appropriate use of inductive techniques.

- The technique is useful if there are documented examples or if they can be obtained easily. It is not suitable where an unpredictable sequence of observations drives the system (e.g. as in some real-time situations).
- The technique is consistent and unbiased and is very suitable for domains where rules form a major part of the knowledge representation.
- Induction provides the knowledge engineer with questions, results and hypotheses which form a basis for consultation with the expert.
- There is no explanation for the rules produced. All output must be examined critically.
- The process assumes that the example set is complete and current.
- Results should not be sensitive to small changes in the training set.

The inductive technique has been used for weather prediction, predicting the behaviour of a new chemical compound, diagnosing plant disease, symbolic integration, improved debt collection, and designing gas-oil separators. See Bratko and Kononenko (1987), Michalski and Chilausky (1980), and Mitchell *et al.* (1983) for examples.

The technique is particularly useful when a great body of data exists about a process but the underlying rules are not known. Induction has been used therefore on large collections of historical process data about industrial plants in order to induce the rules of its operation. Once the rules are known the process can often be optimised. A well-known example of the use of this technique was at the Westinghouse Corporation where over \$10,000,000 was saved (Westinghouse, 1984).

Another interesting use of inductive techniques which will have a wide application in industrial control is its use in conjunction with a qualitative model of the process. This was first carried out in the analysis of electro-cardiograms (Lavrac *et al.*, 1985). A qualitative model of the domain is built. Then, components are failed and the consequences on measurable parameters determined for this failure. The process is repeated for each component and this builds up a complete set of examples of failure. The examples are then used as input to the ID3 algorithm and the rules governing the failure are

induced. These form the basis for a diagnostic expert system. The technique will be discussed in more detail when we examine the application of inductive techniques to satellite power system diagnosis.

BOTTOM-UP OR TOP-DOWN?

There are two competing views about the knowledge acquisition task which might be described as bottom-up and top-down. The bottom-up proponents aim to prise data and concepts out of the expert and then iteratively refine it. Feigenbaum, for example has described knowledge acquisition as "mining those jewels of knowledge out of their (the experts) minds one by one" (Feigenbaum and McCorduck, 1983). The implication is that deeper mining will reveal more relevant knowledge, but this assumes that there is a simple relationship between what is verbalised by experts and what is actually going on in their minds. Hayes-Roth *et al.* (1983) claim that the building of expert systems "is inherently experimental" and is therefore characterised by rapid prototyping which is essentially a bottom-up process. The basic assumption underlying this bottom-up approach is that an expert system is based upon a large body of domain specific knowledge and that there are few general principles underlying the organisation of the domain knowledge in an expert's mind. However, the existence of underlying principles and causal relationships (Davies, 1983) may be an indication that expert knowledge is more domain independent than was assumed by Feigenbaum (1979). Breuker and Wielinga (1987), for example state that "In our experience over the past three years in analysing eight widely different domains a number of concepts have invariably recurred, such as 'procedure', 'process', 'quantification object' . . . and 'identification object' Such concepts are abstractions of real world knowledge". So "expert behaviour that is seemingly domain-specific may originate from higher level problem solving methods which are well-structured and have some degree of domain independence". Domain-independent aspects to the problem solving process have been observed by Pople (1982) in medical diagnosis tasks.

Breuker and Wielinga (1987) strongly support the top-down alternative and claim that there is a crucial step missing in the prototyping approach between the identification of the relevant characteristics of the domain and selection of solution methods, that of "the interpretation of the data into some coherent framework, a model, schema or canonical form". They equate it to the knowledge level of

Newell (1980) or the "missing level" of Brachman (1979) in semantic network analysis. They propose five levels of knowledge analysis—identification, conceptualisation, epistemological, logical and implementational, and have developed these ideas into a knowledge acquisition methodology called KADS [Knowledge Acquisition and Documentation Structuring; Breuker and Wielinga (1985)]. An example of the application of the technique to insurance underwriting is given in Hayward *et al.* (1988).

A COGNITIVE TASK ANALYSIS APPROACH

Roth and Woods (1989) identify "failing to appreciate the demands of the task" as a major reason for the failure in current expert systems developments. They identify the iterative refinement approach (Hayes-Roth *et al.*, 1983) used almost universally during the knowledge acquisition phase as the main cause. From a small prototype, the full system is developed through iterative refinements until the final delivery system is produced. They claim that "the amount of time and resources typically available for systems development in industry does not allow for the long term evolution of systems entailed in the iterative refinement approach" and point out that "architectures which are built based on consideration of a core set of examples will often not have the necessary structural hooks and processing mechanisms to deal with new cases that have complex aspects that had not been represented in the original set" (Bachant and McDermott, 1984). The correct handling of new cases then requires major restructuring of the knowledge rather than fine tuning. Experts often state rules to which there are exceptions, not usually revealed until much later.

They further point out that systems designed from a core set of examples often result in oversimplified representation of goals and constraints and this leads to optimisation of one dimension of the user's problem at the expense of ignoring other goals. One example from process control, given in Roth and Woods (1989) concerned the design of an AI system to support operators in the start-up procedure for a boiler. The AI developers had originally concentrated upon a single goal—that of preventing shut-down. However the operators, in reality, had other goals to meet as well (shut-down could be caused by other sources). Thus, there were circumstances where sub-optimal performance on the boiler level goal was appropriate. They claim that their up-front analysis of the demands of the complete task enabled a much more realistic system to be built (Woods and Roth,

1988). They suggest a multi-phase progression from initial informal interview techniques (to derive a preliminary mapping of the semantics of the domain), to more structured knowledge elicitation techniques (to refine the initial semantic structure), to controlled experiments designed to reveal the knowledge and processing strategies utilised by domain practitioners.

The first phase gives preliminary cognitive description of task to guide further analysis. It is important here not to home in on specific rules. One possibility is to get the experts to provide an overview presentation (Gammack and Young, 1985). Only when an overview of the semantics of the application has been developed can more structured techniques be used.

The second phase concentrates on how practitioners perform their tasks, thus, there is emphasis on observation and analysis of actual task performance. It will involve techniques such as critical incident review, discussion of past challenges, or the construction of test cases on which to observe the experts at work. During this phase, Roth and Woods (1989) also recommended the use of "expert panels" to obtain a corpus of challenging cases to identify critical elements and strategies for handling them.

The third phase uses observational techniques under controlled conditions to observe expert problem solving strategies. The practitioner is observed and asked to provide a verbal commentary. The task can be deliberately manipulated, for example, by forcing the expert to go beyond reasonably routine procedures. In some cases, the expert himself controls the information gathering. Alternatively, it is controlled by the observer. Each approach provides useful information, the former provides data on the diagnostic search process and the latter on the effect (or bias) of particular types of information on expert interpretations. Another useful technique is to compare the performance of experts with different levels of expertise, so as to isolate what factors really account for superior performance.

Roth and Woods (1989) make a strong case for cognitive task analysis approach as compared with a technology-driven approach where knowledge acquisition concentrates upon AI representation mechanisms (e.g. rules and frames).

KNOWLEDGE ACQUISITION TOOLS

A large number of tools for supporting the knowledge acquisition process have been developed in the academic environment and some

of these have been mentioned already. The general aim of all these tools is to minimise the number of iterations needed for the whole knowledge engineering process by bridging the gap between the problem domain and the implementation. Boose and Gaines (1988) give a brief summary of the main tools under development and provide a summary. Some tools endeavour to make the process fully automatic. KRITON (Diederich *et al.*, 1987), for example, has a set of procedures pre-stored—interviews, incremental text analysis, and protocol analysis. Repertory Grids are used to pull out declarative knowledge. An intermediate knowledge representation system is suggested for supporting the knowledge elicitation techniques. The knowledge representation scheme involves a propositional calculus for representing transformations during the problem solving process and a descriptive language for functional and physical objects. This is then translated semi-automatically into the run-time system but this commits the knowledge engineer to a particular representation. Other tools (for example KADS and ACQUIST) merely provide a set of tools to aid a more methodological approach. Thus, KADS aims only to produce a document describing the structure of the problem in the form of a documentation handbook.

KRITON supports only bottom-up knowledge acquisition but KADS supports both top-down and bottom-up approaches. KADS supports bottom-up through a hypertext protocol editor (PE.D) and hierarchies are developed and manipulated by a context editor (CE). Top-down is supported by a set of interpretation models each describing the meta-level structure of a generic task.

The KADS methodology is based upon the following principles:

- knowledge and expertise should be analysed before the design and implementation starts, i.e. before an implementation formalism is chosen
- the analysis should be model driven as early as possible (see also Su, 1988)
- expert problem solving should be expressed as epistemological knowledge
- the analysis should include the functionality of the prospective system
- the analysis should be breadth-first allowing incremental refinement
- new data should only be elicited when previous data has been analysed
- all collected data and interpretations should be documented.

The approach produces a four layer model of expertise (Hayward *et al.*, 1988):

- definition of the domain concepts and their static relationships
- definition of relations arising in a task context which are concerned with dynamics and are expressed in the inference structure
- specification of how the available inferences can be used to undertake a particular task
- definition of how the task level may be controlled. This is the least developed part of the model.

KEATS-1 (Motta *et al.*, 1988) provided a Cross Reference Editing Facility (CREF) and a Graphical Interface System (GIS), to support data analysis and domain conceptualisation. CREF organises the verbal transcript text into segments and collections and GIS allows the knowledge engineer to draw and manipulate domain representations on a sketch pad. In KEATS-2, these have been replaced by ACQUIST, a hypertext application for structuring the knowledge from the raw text data. Fragments from the data are collected around concepts, concepts are factored into groups, and groups into meta-groups. Links can then be defined between any of these entities. The emerging structure is displayed graphically. ACQUIST provides support for both bottom-up approaches (fragments to concepts to groups to meta-groups) and top-down approaches (using what are called coding sheets on which a "caricature of the observed behaviour of the domain expert" is captured). In this approach, the knowledge engineer uses a predefined abstract model to guide the knowledge acquisition process. Use of such models (even if incomplete or inadequate) can dramatically improve the knowledge acquisition process. The coding sheet is a set of hypertext cards.

The knowledge acquisition tool by Strothotte and Sack (1988) is based on the assumption that it is often quite natural for domain experts to express themselves through diagrams. These diagrams and the related dialogue allow the expert to transfer the knowledge in a way often used among humans. The diagrams are drawn with lines by using a simple graphical editor. All details which are important for describing certain objects have to be included. Then, the tool extracts features from the diagram by applying computational geometry and image processing algorithms. Clarifying questions are then asked by the computer to the expert about features which have to be further specified. Knowledge about the objects and their relationships are derived and stored in the final knowledge base

together with the diagram itself. Thus, information content of diagrams can be entered semi-automatically into the knowledge base. Diagrams can be re-used if necessary. Strothotte and Sack stated that their knowledge acquisition tool needs to be combined with a tool for textual knowledge in all those domains which allow to describe only some types of knowledge in a diagrammatic way. A further limitation may be that the expert will be forced to overspecify irrelevant details.

A further knowledge acquisition tool is ROGET (Bennett, 1985). It conducts a dialogue with a domain expert in order to acquire his or her conceptual structure. ROGET gives advice on the basis of abstract categories and evidence. Initial conceptual structures are selected on this basis. Only a small set of example systems were tested.

The use of Pathfinder networks for knowledge acquisition was proposed by Esposito and Dearholt (1988). It is a tool for the identification of conceptual structures with a sophisticated interactive graphics system for network display and manipulation. Experts are asked to make simple similarity judgements and answer specific questions. This graph-theoretic tool has been used in investigations with network models of human semantic memory utilising estimates of psychological distance. Path Algebra techniques (Alty and Richie, 1985) provide a more generalised tool for such approaches.

The systematic acquisition of knowledge about the fault behaviour of a technical system was suggested by Narayanan and Viswanadham (1987). A procedure involves the development of a hierarchical failure model with fault propagation digraphs and cause-consequence knowledge bases for a given system. It uses the so-called augmented fault tree as an intermediate knowledge representation. Fault propagation digraphs describe the hierarchical structure of the system with respect to faults in terms of propagation. The cause-consequence knowledge bases characterise failures of subsystems dependent on basic faults by means of production rules. The knowledge acquisition process can be reduced to defining parameters required by the knowledge representation scheme and transforming human expertise into these parameter values. The augmented fault tree is a conceptual structure, which describes causal aspects of failures as in conventional fault trees but additionally also probabilistic, temporal and heuristic information. The production rules of cause-consequence relations are derived from the augmented fault tree by decomposing it into mini fault trees. The proposed methodology

has reached a relatively high level of formal description. However, it cannot yet deal with inexact knowledge by using ranges of parameters. An example of a failure event in a reactor system is given.

ELICITATION EXAMPLES

The application of the knowledge elicitation techniques for industrial expert systems will be shown with two examples. The first is the task and knowledge analysis in power plants performed with using observation interviews, questionnaires and time lines, and the second is the knowledge and task analysis performed in parallel to the construction of a failure management expert system for space systems.

Knowledge analysis in power plants

An extensive task and knowledge analysis has been performed in coal-fired power plants and in a power plant school by the first author of this paper and his research group (Borys *et al.*, 1987; Johannsen *et al.*, 1987; Johannsen, 1989; Sundström, 1990). The work is part of the ESPRIT-GRADIENT project on "Graphics and Knowledge Based Dialogue for Dynamic Systems" which is partially supported by the Commission of the European Communities and is performed in cooperation between the research groups of the two authors of this paper together with two industrial companies from Germany and Denmark and a Belgian university.

A thermal power plant consists basically of a water-steam cycle involving a boiler, turbines, condenser and feedwater system, as well as a generator. The automation system or supervision and control of the plant is hierarchically organised into drive, group and control levels. When higher levels of the automation system fail, the shift leader needs to operate the plant with less automation on the lower levels. It is intended to support the human operators in these situations using expert systems. Several cooperative expert systems are developed within the whole research consortium, mainly for diagnosis of causes of failures, prediction of consequences of failures, knowledge-based alarm handling, procedural support of operator behaviour and plan recognition with operator input evaluation. Also, a graphical expert system will be developed. It will be based on intelligent graphical editors which contain knowledge-based support functions for graphical picture designers.

The task and knowledge analyses were performed during a period of three years in several power plants during day and night shifts

as well as in a power plant school. Observation interviews, questionnaires and time lines were used as elicitation techniques. During later stages of the elicitation process, the analyses were restricted to failure situations in the pre-heating system. The knowledge was collected for two reasons—to build a diagnostic expert system for supporting operators and to construct a user model which will form part of an intelligent graphical editor to support designers. After using interviews and observations with power plant operators and operational engineers, a number of different questionnaires were applied. A state-oriented questionnaire with a total number of 29 questions helped to build frames of knowledge for each substate of the plant. The questionnaire was structured into the five groups of substate description, activities, mental models of the operator, effects of activities, and suggestions for improvement with respect to the operator's work. Another questionnaire was designed to capture expert strategies and knowledge representations available to operators in failure situations. This failure-oriented questionnaire was based on a general separation of each failure situation into several phases of fault management such as detection, diagnosis, localisation, compensation and correction (Johannsen, 1988).

Time lines were used in later stages of the task and knowledge analyses. They were mainly applied in the power plant school where it is possible to freeze a system state of interest in the simulator, measure the time of occurrence and collect the related knowledge from all available sources without any pressure. For the diagnostic expert system, knowledge collected and formalised using the time lines approach includes alarm messages, affected components, parameters, actions of the supervision and control system, and human operator actions. The knowledge elicitation for the user model of the intelligent graphical editor is concerned with decision alternatives and associated information search behaviour of the operator (Sundström, 1990). It is related to the information processing goals of categorisation of states, choice of actions, and evaluation of outcomes. This information was also elicited using time lines. The information gathered in each of the two types of time line is different but both types are related to each other with respect to time for the same failure situation. Each of these time lines can be viewed as a kind of intermediate knowledge representation on paper. This technique is a useful tool for further knowledge formalisation and for knowledge implementation. Time lines can easily be discussed with domain experts before

any implementation needs to be accomplished. In the case of this project, the time lines, as well as first prototype implementations, were evaluated by a power plant instructor (the domain expert) together with two researchers (the knowledge engineers).

The effort for the knowledge elicitation in the task and knowledge analyses was approximately the same for both systems, the diagnostic expert system as well as the user model. The first phase in the sense of a cognitive task analysis approach as mentioned above with initial informal interviews took half a month each for both systems. The later two phases with more detailed analyses required an effort of 5 months for the diagnostic expert system and almost 4 months for the user model. This included 2 full working days each for both systems at the simulator in the power plant school as well as a total of 14 days of sessions with experts for the diagnostic expert system and 8 days for the user model.

The last prototype version of the diagnostic expert system contains knowledge bases with about 370 units (objects) as well as about 100 rules in 14 rule classes. The branching factor is 3–14 for each rule and the chaining factor 4–7. The number of units of the user model prototype is about 170 and the number of rules is 41 in 4 rule classes. 84 states can be differentiated in the user model.

Construction of failure management expert systems

Another example of the application of knowledge elicitation techniques was given in the construction of failure management expert systems by Malin and Lance (1987). A knowledge and task analysis was performed in parallel with the construction of an expert system for failure management in a space station prototype device. A device for removing carbon dioxide from cabin air was selected. The expert system is called FIXER (Fault Isolation Expert to Enhance Reliability).

The knowledge engineering process for developing FIXER was performed by three persons: an expert in life support systems, a cognitive scientist and a consulting knowledge engineer. The prototype was developed through close cooperation between the systems expert and the cognitive psychologist who was the main knowledge engineer. The knowledge for the trouble-shooting expert system was based on device design information and experience with similar devices. Operational knowledge about trouble-shooting failures in the device was not used. The goals, tasks, knowledge, methods and

design decisions involved in constructing the failure management expert system were intensively observed and analysed. Thus, observations were the selected knowledge acquisition technique for the investigation of the whole design process. Five design tasks were observed and analysed:

- allocation of failure management functions and interface definition
- analysis of failure events, fault modes, and effects
- selection and construction of measurements and test procedures
- analysis of fault-symptom patterns and construction of diagnosis procedures
- construction of procedures for failure effects management and maintenance.

Furthermore, tasks for the revision of the failure management software were observed and evaluated. The development and analysis effort for FIXER required about 30 full working days for each person, the domain expert as well as the knowledge engineer, distributed over a time of five months. The software revisions were performed one year later and required half a day for the expert and seven days for the knowledge engineer. The FIXER knowledge base contains 198 units and references 42 LISP functions. For diagnosis, 28 rules in 9 rule classes are used.

One of the conclusions drawn from the experience with this cooperative knowledge elicitation technique is that "the knowledge acquired from the expert should include much more than the rules and procedures for failure management". The expert's choice of strategies, and supporting analyses and models, need also to be represented explicitly in order to deal with limitations of the failure management expert system and the need for later revisions. Further, it was observed during the whole design process that mental models of the device and its behaviour were important. Such models are under development by Sundström and Johannsen (1989).

INDUCTION EXAMPLES

Two examples will be given of the use of inductive techniques in the industrial environment—their use in British Petroleum in the design of gas-oil separator plant, and at the European Space Agency for the design of expert systems in satellites.

Use of induction by engineers

The gas-oil system assists engineers to design gas-oil separators. The underlying hydrocarbon production separation process is quite complicated, relying on a variety of knowledge sources such as manuals, codes of practice, space limitations, and on the crude oil quantity and the gas quality required. Key factors include the delivery system, a user friendly interface involving graphics design, and interfacing to existing FORTRAN routines. Gas-oil is a large system—containing over 2500 rules and is expected to eventually grow in size to 100,000 rules. In the present example, the system used 1600 examples to create an expert system occupying about 3 Mbytes of storage. This consisted of 14,500 lines of FORTRAN code generated automatically by the induction process. The automatic generation of FORTRAN was particularly useful since this enabled the system to be interfaced easily to other engineering modules. The remarkable fact about the system is that it took one year of effort to create and now requires about one month per year of maintenance effort. This should be compared with the effort required to produce the MYCIN and XCON systems (100 and 180 man-years, respectively). Much of this reduced development and maintenance cost is claimed to result from the use of inductive techniques (Gurtoyle, 1986).

Slocombe of British Petroleum (Slocombe *et al.*, 1986) claims that the inductive technique is ideally suited to the engineering temperament. "The expert is invited to suggest a possible set of solutions to a particular problem. Then, he thinks of the factors which are involved in deciding which choice to make. For example, in choosing a type of vessel for use in a refinery the expert would take into account the quality of output needed, the throughput of material, the size of the site available and so on. The last stage of this process involves the expert providing a few examples of real cases. The software then induces the rule. At this stage, the expert's interest is caught. The rule may be over simple, so the expert thinks up another example to illustrate the difference between the two cases. Or the rule may separate two factors which are later decided to be the same. This technique concentrates on: homing in on the psychological problems".

The key point is that the domain experts are able to interact directly with the knowledge acquisition process with minimum assistance from a knowledge engineer. As Moore points out (Slocombe *et al.*, 1986), "The academic approach . . . requires a knowledge engineer with

no expertise in the engineering domain to approach the expert and say 'I know nothing about your area, but would like you to tell me everything you know, and preferably in the form of explicit rules'. Our best experts are long experienced individuals with a wealth of heuristics and rules of thumb. They find it very difficult to articulate their knowledge explicitly but can reel off any number of examples and outcomes under particular circumstances". BP combine this inductive approach with sophisticated system building facilities and are now using the inductive techniques in a number of application areas.

Satellite diagnosis

The second application involves the construction of a prototype on-board expert system for dealing with power failures in a satellite. The approach is interesting in that it not only uses qualitative modelling and inductive techniques, but also tries to address the problems of validation mentioned earlier. This latter problem was, in fact, the main reason behind the project since the European Space Agency was naturally very wary of placing unvalidated expert systems on board a satellite without validation. The work was carried out at the Turing Institute, Glasgow, and has been reported by Pearce (1988). The power subsystem for the satellite can be viewed as in Fig. 2.

When the satellite is in sunlight, power is generated in the solar array panels. This power is used to drive the payload and recharge the battery. In the eclipse phase, the battery maintains the payload. The Array Switch Regulator (ASR) contains switches to enable or disable solar panels. A comparator detects a rise or fall in bus voltage and automatically opens or closes ASR switches to restore the bus voltage. The Electrical Integration Unit contains switches for main charging or trickle charging the two batteries. A real-time numerical simulator had been developed for the satellite to be used for operator training. This simulator enabled the

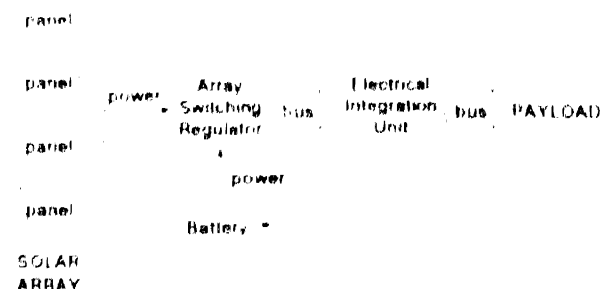


Fig. 2. The satellite power system

developments to be tested on real data. A traditional expert system had already been constructed to perform the above tasks but validation for this system was problematic. The approach was to construct an expert system in such a way that the validation problem was minimised.

The solution involved three key components—a qualitative model of the system, the use of rule induction and a sophisticated display system to make the operation of the satellite visible. The approach is quite general and can be applied to most physical systems.

The qualitative model of the satellite power system was constructed in the language PROLOG. It is a model containing deep causal knowledge as opposed to shallow, operational knowledge. Shallow knowledge is sufficient to perform a task but does not have the underlying causal mechanisms. Deep knowledge, on the other hand, allows reasoning from first principles to be carried out. The model consists of definitions of each component and its relationship to other components (upstream and downstream of it). Behaviour rules specify how each component affects any neighbouring components. All that is then needed is an initial state and the model can be executed. The model is a qualitative one in that it does not deal with precise values of voltages or currents over time. Rather it deals with values such as low, medium and high. Although, at first sight, this seems very restrictive, it corresponds closely to domain expert descriptions when reasoning about the operation of devices, and it is able to give more meaningful explanations than a quantitative model. Of course, it is also computationally less complex.

The satellite transmits a number of parameter values to the ground control and the mechanism for inducing a failure rule is as follows:

- fail a component
- run the model until it reaches a steady state
- collect the values of the observable parameters
- the failure state together with the observable set make up an example of failure
- repeat this for all single components; this results in a complete set of failure examples
- use rule induction to induce the underlying rules of failure
- the rule is added to the expert system.

The process thus generates shallow knowledge from deep knowledge with a considerable saving in storage. For example, the deep knowledge representation required 36 Kbytes of store whereas the shallow representation required only 9. Although the validation problem has not

been completely solved (is it actually solvable?) the reliability of the rule system has been improved. Since the failure set is a complete set the validation process has been simplified to validation of the qualitative model and this is much easier to do. The model when taken in conjunction with the graphical user interface enables the engineer to visually check the model. This is not possible with a set of rules.

To enable the reliability of the rule set to be checked and compared with the original expert system, a knowledge integrity checker was constructed which checks a rule set for some key features—unreachable clauses, dead-end clauses, cyclic clauses, type checking, incompleteness and subsumption (Pearce, 1987). The checker revealed that the original rule set (which consisted of 110 rules) contained seven errors—a type error, two unreachable clauses, and four dead-end clauses. The induced rule set (75 rules) had no errors. Finally, both expert systems were executed on the simulator. Fourteen different error conditions were simulated. The original system detected 10 of these (72%). The induced rule set was 100% successful. One other important aspect to note is that the development time of the original expert system was 6 man-months and the total development time for the induced set (including the building of the qualitative model) was 3–4 man-months.

The bottleneck which still exists in using this approach is that required for the creation of the PROLOG model. The Turing Institute is currently automating this process by using SGML (Standardised Generalised Mark-up Language) (Smith and Stutely, 1988) for specifying the model. The model is specified in normal text using SGML markers to identify key model components and interactions between them. The PROLOG model is then automatically generated from the SGML document using a translator.

This example shows the power of the approach. Furthermore, it can deal with multiple error situations—one simply creates examples of multiple error situations. Although this is computationally intensive it only has to be done once. For complex systems, the system can be broken down into lower level subsets and solved individually. Then, the interactions between these subsets can be treated separately. The approach is ideally suited to diagnostic problems in industrial systems where validation is important. Finally if the deep model is maintained along with the shallow model, excellent explanations can be provided.

CONCLUSIONS

Kidd (1987) has given the following guidelines to assist in the knowledge acquisition process:

- Knowledge engineers should appreciate that any data elicited have to be interpreted as to what underlying knowledge or processes they imply and that the system knowledge base is a model of the expert's domain knowledge constructed by the knowledge engineer.
- The nature and difficulty of the knowledge acquisition process depends on the degree of formality of the human language used to describe the domain.
- Knowledge engineers should aim to formulate a knowledge level description of the selected problem solving task.
- At an early stage, knowledge engineers need to decide on the appropriate modality for the proposed system.
- In any interactive system, the user is the active agent. Analysing user requirements must be a key part of the knowledge acquisition.
- If the role is decision support, the knowledge engineer should identify weaknesses in the expert reasoning process and aim to complement these.

These guidelines summarise what has been outlined and emphasised in the previous sections of this paper. The importance of differentiating between the task and the implementation level has been highlighted. It is clear that the cognitive behaviour and knowledge structures in human problem solving tasks have to be understood and formalised by appropriate knowledge acquisition techniques, before any systems implementation makes sense. These factors are not well understood at the present time. Equally importantly, the knowledge elicited from a domain expert must be checked and re-checked for possible cognitive bias.

Because the only known way to test a knowledge base is to execute it, acquisition and implementation often become closely intertwined. This increases the danger that the technology used for implementation will influence the acquisition process itself. One way of minimising this danger is to use some form of intermediate knowledge representation formalism or language (for example Matsumura *et al.*, 1989), however, we are still some way from achieving this objective. This fact, together with the requirement for several iterations over the knowledge acquisition process will inevitably mean that knowledge engineering will be a skilled and time consuming activity for the foreseeable future.

There is clearly an urgent need for improved tools. More tools should be designed which allow domain experts to play a more direct role in the knowledge acquisition activity. There is also a need for tools to assist with the validation

problem, a crucial area in the development of industrial expert systems. In this respect, the qualitative modelling approach (when coupled with inductive techniques) has shown some promise in at least pushing back the validation process away from direct rule validation towards model validation which is somewhat easier to do. In the absence of formal validation, testing will continue to play a major role.

Some workers (for example Krasemann and Krasemann, 1988) have emphasised the need for a thorough systems analysis approach (as used in software engineering) in the development of expert systems. Whilst some artificial intelligence techniques are beginning to be used in software engineering (for example selection of conceptual structures and knowledge representation models and the use of rapid prototyping), there are proven procedures from conventional software engineering which could be used to advantage in the knowledge engineering process. Such procedures would include requirements analysis, testing strategies, approaches to realising efficient implementations and maintenance/revision techniques. A requirement and function analysis needs to be performed. The criteria for testing the system have to be derived from the requirement analysis. Software optimisation, portability between languages and use of tools with efficient run-time environments are all important. Ideally, Krasemann and Krasemann (1988) suggest that the software maintenance should be left with the system user and domain expert rather than with the knowledge engineer.

Can the domain experts also take on the role of the knowledge engineer? It is certainly true that the effort required for a domain expert to master the techniques of knowledge engineering is less than that required for the knowledge engineer to become a domain expert. Certainly, this view is more plausible in the field of industrial expert systems than in other domains. Industrial domain experts are better trained to think in systems and software engineering terms. Therefore, they should more easily be able to express their knowledge in paradigms which are more closely related to system implementation provided appropriate knowledge acquisition tools exist. An example in this direction has been given above in the construction of the failure management expert system. However, experts by their very nature are scarce and overworked, so it is unlikely that they will have the motivation to do this. There will also be difficulties in eliminating cognitive bias. On the whole, we do not think that this is a practical solution, particularly with the relatively poor state of current knowledge acquisition tools.

Another problem which has not yet been

solved concerns the conflict between the different depths of knowledge required by users with different levels of expertise and skill. The knowledge requirements for expert systems to be used by other high-level experts will differ from those aimed at advising less expert or casual users (Madni, 1988). This conflict has not been properly recognised or fully investigated. The construction of expert systems which support different kinds of users with different degrees of expertise and knowledge will require a much deeper understanding of the cognitive behaviour and knowledge structures used in human problem solving tasks.

Acknowledgements—This work was partially supported by the Commission of the European Communities through the ESPRIT-GRADIENT project P857. The project is carried out by a consortium consisting of Axion (Copenhagen, Denmark), Asea Brown Boveri (Heidelberg, F.R.G.), University of Kassel (Kassel, F.R.G.), University of Strathclyde (Glasgow, Scotland, U.K.), and the Catholic University of Leuven (Leuven, Belgium).

REFERENCES

- Alty, J. L. and G. Johannsen (1989). Knowledge based dialogue for dynamic systems. *Automatica* **25**, 829-840. [Also in R. Isermann (Ed.) *Proc. IFAC 10th World Congress on Automatic Control* (Preprints, 1987, **7**, 358-367)].
- Alty, J. L. and R. Richie (1985). A path algebra facility for interactive dialogue designers. In P. Johnson and S. Cook (Eds.), *People and Computers. Designing the User Interface*. Cambridge University Press, Cambridge, pp. 128-137.
- Bachant, J. A. and J. McDermott (1984). RI revisited: Four years in the trenches. *The AI Magazine*, **4**, 21-32.
- Bennett, J. S. (1985). ROGEE: A knowledge-based system for acquiring the conceptual structure of a diagnostic expert system. *J. Automated Reasoning*, **1**, 49-74.
- Boose, J. H. (1986). *Expertise Transfer for Expert Systems Design*. Elsevier, New York.
- Boose, J. H. and J. M. Bradshaw (1988). Expertise transfer and complex problems. Using AQUINAS as a knowledge acquisition workbench for knowledge-based systems. *Int. J. Man-Machine Studies*, **2**, 39-64.
- Boose, J. H. and B. R. Gaines (1988). Knowledge acquisition tools for expert systems. In B. R. Gaines and J. H. Boose (Eds.), *Knowledge Acquisition Tools for Expert Systems*, Vol. 2. Academic Press, London, pp. xiii-xvi.
- Borys, B.-B. (1988). Ways of supporting ergonomically and technically correct display design. *Proc. Workshop Human Computer Interaction and Complex Systems*. Alexandria, Scotland.
- Borys, B.-B., G. Johannsen, H.-G. Hansel and J. Schmidt (1987). Task and knowledge analysis in coal-fired power plants. *IEEE Control Syst. Magazine*, **7**, 26-30.
- Brachman, R. J. (1979). *A structured paradigm for representing knowledge*. BBN Technical Report, Bolt, Beranek and Newman, Cambridge, MA.
- Bratko, I. (1989). Machine learning. In K. J. Gilhooly (Ed.), *Human and Machine Problem Solving*. Plenum Press, New York, pp. 265-286.
- Bratko, I. and I. Kononenko (1987). Learning diagnostic rules from incomplete and noisy data. In B. Phelps (Ed.), *Interactions in AI and Statistics*. Gower Technical Press, London.
- Breuker, J. A. and B. Wiehinga (1985). KADS: Structured knowledge acquisition for expert systems. *Proc. 5th Int. Workshop Expert Systems and their Applications*. Avignon.
- Breuker, J. A. and B. Wiehinga (1987). Use of models in the interpretation of verbal data. In A. L. Kidd (Ed.), *Knowledge Acquisition for Expert Systems*. Plenum Press, New York, pp. 17-44.
- Buchanan, B. G., D. Barstow, R. Bechtal, J. Bennett, W. Clancey, C. Kulikowski, T. Mitchell and D. A. Waterman (1983). Constructing an expert system. In F. Hayes-Roth, D. A. Waterman and D. B. Lenat (Eds.), *Building Expert Systems*. Addison-Wesley, Reading, MA, Chapter 5.
- Chignell, M. H. and J. G. Peterson (1988). Strategic issues in knowledge engineering. *Human Factors*, **30**, 381-394.
- Clancey, W. J. (1986). Transcript of plenary sessions: Cognition and expertise. *1st AAAI Workshop Knowledge Acquisition in Knowledge Based Systems*. Banff, Canada.
- Krasemann, C. and H. Krasemann (1988). Der Wissens-Ingenieur—ein neuer Hut auf altem Kopf. *Informatik-Spektrum*, **11**, 43-48.
- Davies, R. (1982). *Expert Systems: Where are we? and Where do we go from here?* AI Memo No. 665, MIT AI Laboratory.
- Davies, R. (1983). Reasoning from first principles in electronic trouble shooting. *Int. J. Man-Machine Studies*, **19**, 403-423.
- DeMontmollin, M. and V. DeKeyser (1986). Expert logic versus operator logic. In G. Johannsen, G. Mancini and L. Mårtensson (Eds.), *Analysis, Design, and Evaluation of Man-Machine Systems (Proc. 2nd IFAC/IFIP/IFORS/IEA Conf.)*. Pergamon Press, Oxford, pp. 43-49.
- Diederich, J. and M. Linster (1989). Knowledge-based knowledge elicitation. In G. Gorda and C. Tasso (Eds.), *Topics in Expert Systems Design*. North Holland, Amsterdam, p. 325.
- Diederich, J., I. Ruhmann and M. May (1987). KRITON: a knowledge acquisition tool for expert systems. *Int. J. Man-Machine Studies*, **26**.
- Elzer, P. and G. Johannsen (Eds) (1988). *Concepts, Design, and Prototype Implementations for an Intelligent Graphical Editor (IGE 1)*. ESPRIT-GRADIENT P857, Report No. 6. Labor. Man-Machine Systems, University of Kassel (GhK).
- Eriksen, K. and H. A. Simon (1984). *Protocol analysis. Verbal Reports as Data*. MIT Press, Cambridge, MA.
- Esposito, C. and D. Dearholt (1988). Pathfinder networks for knowledge acquisition: Cognitive structures, network representations, and display. *Preprints 3rd IFAC/IFIP/IEA/IFORS Conf. Man-Machine Systems*. Oulu, Vol. 1, pp. 76-81.
- Feigenbaum, E. A. (1979). Themes and case studies in knowledge engineering. In D. Michie (Ed.), *Expert Systems in the Microelectronic Age*. Edinburgh University Press, Edinburgh.
- Feigenbaum, E. A. and P. McCorduck (1983). *The Fifth Generation*. Addison-Wesley, New York.
- Forsythe, D. E. and B. G. Buchanan (1989). Knowledge acquisition for expert systems. Some pitfalls and suggestions. *IEEE Trans. Syst. Man Cybern.*, **19**, 435-442.
- Fox, J., C. D. Myers, M. F. Greaves and S. Pegram (1987). A systematic study of knowledge base refinement in the diagnosis of Leukemia. In A. L. Kidd (Ed.), *Knowledge Acquisition for Expert Systems*. Plenum Press, New York, pp. 73-90.
- Gammack, J. G. and R. M. Young (1985). Psychological techniques for eliciting expert knowledge. In M. Bramer (Ed.), *Research and Development in Expert Systems*. Cambridge University Press, Cambridge, U.K.
- Gruber, T. R. and P. R. Cohen (1987). Design for acquisition: Principles of knowledge system design to facilitate knowledge acquisition. *Int. J. Man-Machine Studies*, **26**, 143-159.
- Guilloyle, C. (1986). Ten minutes to lay the foundations. *Expert System User*, August 16-19.
- Harmon, P. (1989). Thirty successful systems. *Expert Systems Strategies*, **5**, Cutter Information Corp., Arlington, U.S.A.
- Hart, A. (1987). Role of induction in knowledge elicitation. In A. L. Kidd (Ed.), *Knowledge Acquisition for Expert Systems*. Plenum Press, New York, pp. 165-189.

- Hayes-Roth, F., D. A. Waterman and D. B. Lenat (1983). *Building Expert Systems*. Addison-Wesley, Reading, MA.
- Hayward, S. A., B. Wielinga and J. A. Breuker (1988). Structured analysis of knowledge. In J. H. Bouwe and B. R. Gaines (Eds), *Knowledge Acquisition Tools for Expert Systems*. Academic Press, London, pp. 140-160.
- Hollnagel, E. (1987). Plan recognition in modelling of users. *Reliability Engineering and System Safety*, **22**, 129-136.
- Jackson, P. (1985). Reasoning about belief in the context of advice-giving systems. In M. Bramer (Ed.), *Research and Development in Expert Systems*. Cambridge University Press, Cambridge, U.K.
- Johannsen, G. (1988). Categories of human operator behaviour in fault management situations. In I. P. Goodstein, H. B. Andersen and S. E. Olsen (Eds), *Task, Errors and Mental Models*. Taylor & Francis, London, pp. 251-258.
- Johannsen, G. (1989). Knowledge analysis in power plants. In M. G. Singh (Ed.), *Systems and Control Encyclopedia First Supplement*. Pergamon Press, Oxford, pp. 366-373.
- Johannsen, G. (1990). Towards a new quality of automation in complex man-machine systems. *Preprints IFAC 11th World Congress Automatic Control*, Tallinn, Estonia, USSR, Vol. 10, pp. 127-181.
- Johannsen, G., S. Borndorff and G. A. Sundstrom (1987). Knowledge elicitation and representation for supporting power plant operators and designers. *Proc. 1st European Meeting Cognitive Science Approaches to Process Control*, Marcoussis, France, pp. 1-10.
- Johnson, L. E. and N. E. Johnson (1987). Knowledge elicitation involving teachback interviewing. In A. E. Kidd (Ed.), *Knowledge Acquisition for Expert Systems: A Practical Handbook*. Plenum Press, New York, pp. 91-108.
- Kelly, G. A. (1955). *The Psychology of Personal Constructs*. Norton, New York.
- Klein, G. A., R. Calderwood and D. MacGregor (1989). Critical decision method for eliciting knowledge. *IEEE Trans. Syst. Man Cybern.*, **19**, 462-472.
- Kidd, A. E. (1987). *Knowledge Acquisition for Expert Systems*. Plenum Press, New York.
- Kuijpers, B. and J. P. Kassirer (1987). Knowledge acquisition by analysis of verbal protocols. In A. E. Kidd (Ed.), *Knowledge Acquisition for Expert Systems*. Plenum Press, New York, pp. 45-71.
- Lavrac, N., I. Brajko, I. Mozatec, B. Cercek, M. Horvat and D. Grad (1985). KARDIO-E: An expert system for electrocardiographic diagnosis of cardiac arrhythmias. *Expert Systems*, **2**, 46-50.
- MacDonald, B. A. and I. H. Witten (1989). A framework for knowledge acquisition through techniques of concept learning. *IEEE Trans. Syst. Man Cybern.*, **19**, 499-512.
- Madni, A. M. (1988). The role of human factors in expert systems design and acceptance. *Human Factors*, **30**, 395-414.
- Malin, J. E. and N. Lance (1987). Processes in construction of failure management expert systems from device design information. *IEEE Trans. Syst. Man Cybern.*, **SMC-17**, 956-967.
- Matern, B. (1984). *Psychologische Arbeitsanalyse*. Springer, Berlin.
- Matsumur, S., K. Kawai, A. Kamiya, T. Koi and K. Momoeda (1989). User-friendly expert system for turbine/generator vibration diagnosis. *Proc. 4th IFAC/IFIP/IFORS/IEA Conf. Man-Machine Systems*, Xian, China, pp. 63-68.
- Michalski, R. S. and R. L. Chilauskys (1980). Learning by being told and learning from examples: An experimental comparison of two methods for knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *Int. J. Policy Anal. Inform. Syst.*, **4**, 125-161.
- Michie, D. and R. Johnston (1985). *The Creative Computer*. Pelican, London.
- Mitchell, T. M., P. E. Utgoff and R. Banerji (1983). Learning by experimentation: Acquiring and refining problem solving heuristics. In R. S. Michalski, J. Carbonell and T. M. Mitchell (Eds), *Machine Learning: An Artificial Intelligence Approach*. Tioga, Palo Alto.
- Motta, E., M. Eisenstadt, K. Pitman and M. West (1988). Knowledge acquisition in KEATS: The knowledge engineers assistant. *Expert Systems*, **5**.
- Motta, E., T. Rajan and M. Eisenstadt (1989). Knowledge acquisition in KEATS-2. In G. Guida and C. Tasso (Eds), *Topics in Expert Systems Design*. North Holland, Amsterdam, p. 299.
- Mullin, I. M. (1989). Experts estimation of uncertain quantities. *IEEE Trans. Syst. Man Cybern.*, **19**, 616-625.
- Narayanan, N. H. and N. Viswanadham (1987). A methodology for knowledge acquisition and reasoning in failure analysis of systems. *IEEE Trans. Syst. Man Cybern.*, **SMC-17**, 274-288.
- Newell, A. (1980). Physical symbol systems. *Cognitive Science*, **4**, 333-361.
- Newell, A. and H. A. Simon (1972). *Human Problem Solving*. Prentice Hall, Englewood Cliffs, NJ.
- Pask, G. (1974). *Conversation, Cognition and Learning: A Cybernetic Theory and Methodology*. Elsevier, London.
- Pearce, D. A. (1987). *KIC: A Knowledge Integrity Checker*. Turing Institute Report No. TIRM-87-025 (available from the Turing Institute, Glasgow, Scotland, U.K.).
- Pearce, D. A. (1988). *The Induction of Fault Diagnosis Systems from Qualitative Models*. Turing Institute Report No. TIRM-88-029 (available from the Turing Institute, Glasgow, Scotland, U.K.).
- Pople, H. E. (1982). Heuristic methods for imposing structure on illstructured problems: The structure of medical diagnosis. In P. Szolovits (Ed.), *Artificial Intelligence in Medicine*. Westview Press, Boulder, Colorado.
- Quinlan, R. (1979). Discovering rules from large collections of examples: A case study. In D. Michie (Ed.), *Expert Systems in the Microelectronic Age*. Edinburgh University Press, Edinburgh, Scotland, U.K.
- Rasmussen, J. (1984). Strategies for state identification and diagnosis in supervisory control tasks, and design of computer based support systems. In W. B. Rouse (Ed.), *Advances in Man-Machine Systems Research*, Vol. 1. JAI Press, Greenwich, CN, pp. 139-193.
- Roth, E. M. and D. D. Woods (1989). Cognitive task analysis: An approach to knowledge acquisition for intelligent system design. In G. Guida and C. Tasso (Eds), *Topics in Expert Systems Design*. North Holland, Amsterdam, pp. 233-264.
- Rouse, W. B. (1986). On the value of information in system design: A framework for understanding and aiding designers. *Inform. Process. Management*, **22**, 217-228.
- Rouse, W. B., E. M. Hammer and C. M. Lewis (1989). On capturing human skills and knowledge: Algorithmic approaches to model identification. *IEEE Trans. Syst. Man Cybern.*, **19**, 558-573.
- Shalin, V. L., D. E. Perschbacher and P. G. Jamar (1988). Intent recognition: An emerging technology. *Proc. Int. Conf. Human Machine Interaction and Artificial Intelligence in Aeronautics and Space*, Toulouse, pp. 139-149.
- Shaw, M. E. G. and B. R. Gaines (1986). Techniques for knowledge acquisition and transfer. *Proc. Knowledge Acquisition for Knowledge Based Systems Workshop*, Banff, Canada.
- Shirley, R. S. (1987). Some lessons learned using expert systems for process control. *IEEE Control Syst. Magazine*, **7**, 11-15.
- Slucombe, S. K., K. D. M. Moore and M. Zeloud (1986). *Engineering Expert System Applications*. Offprint (available from the Turing Institute, Glasgow, Scotland, U.K.).
- Smith, J. M. and R. Stutely (1988). *SGMI: The Users Guide to ISO 8879*. Ellis Horwood, Chichester, U.K.
- Strothotte, T. and J. R. Sack (1988). Knowledge acquisition using diagrams. *Preprints 3rd IFAC/IFIP/IEA/IFORS Conf. Man-Machine Systems*, Oulu, Finland, Vol. 1, pp. 69-75.
- Su, S. Q. (1988). The modelling of human expertise in knowledge acquisition process. *Preprints 3rd IFAC/*

- IFIP/IEA/IFORS Conf. Man-Machine Systems*. Oulu, Finland, Vol. 1, pp. 87-91.
- Sundström, G. A. (1988). User modelling: A new technique to support designers of graphical support systems in conventional power plants. *Preprints 3rd IFAC/IFIP/IEA/IFORS Conf. Man-Machine Systems*. Oulu, Finland, Vol. 1, pp. 36-39.
- Sundström, G. A. (1990). User modelling as a method for supporting designers of graphical interfaces. *Interacting with Computers* (to appear).
- Sundström, G. A. and G. Johannsen (1989). Functional information search: A framework for knowledge elicitation and representation for graphical support systems. *Proc. 2nd European Meeting Cognitive Science Approaches to Process Control*. Siena, Italy, pp. 129-140.
- Tolcott, M. A., F. F. Marvin and P. E. Lehner (1989). Expert decision making in evolving situations. *IEEE Trans. Syst. Man Cybern.*, **19**, 606-615.
- Westinghouse (1984). Press Release.
- Woods, D. D. and E. M. Roth (1988). Aiding human performance: II From cognitive analysis to support systems. *Le Travail Humain*, **51**, 139-171.

Decision Trees and Transient Stability of Electric Power Systems*

L. WEHENKEL†‡ and M. PAVELLA†§

A general inductive inference method is proposed and applied to the automatic building of decision trees for the transient stability assessment of power systems. On the basis of large sets of simulations, the essential features of the method are analysed and illustrated.

Key Words—Artificial intelligence, pattern recognition, decision trees, electric power systems, data reduction and analysis, decision theory, stability, transient stability analysis, sensitivity and control.

Abstract—An inductive inference method for the automatic building of decision trees is investigated. Among its various tasks, the splitting and the stop splitting criteria successively applied to the nodes of a grown tree, are found to play a crucial role on its overall shape and performances. The application of this general method to transient stability is systematically explored. Parameters related to the stop splitting criterion, to the learning set and to the tree feature are thus considered, and their influence on the tree feature is scrutinized. Evaluation criteria appropriate to assess accuracy are also compared. Various trade-offs are further examined, such as complexity vs. number of classes, or misclassification rate vs. type of misclassification error. Possible uses of the trees are also envisaged. Computational issues relating to the building and the use of trees are finally discussed.

1. INTRODUCTION

THE DECISION TREE methodology is nowadays recognized to be a generally nonparametric technique, able to produce classifiers in order to assess new, unseen situations, or to uncover the mechanisms driving a problem (Breiman *et al.*, 1984; Friedman, 1977; Kononenko *et al.*, 1984; Quinlan, 1986). The building of a decision tree is based on a learning set (LS), composed of a number of states together with their corresponding known classification. The building procedure starts at the top node of the tree with the entire LS, and progresses by recursively creating successor nodes, i.e. by splitting the LS into subsets of increasing classification purity. The procedure is stopped when all the newly created

nodes are "terminal" ones, containing "pure enough" learning subsets. The ways of splitting the successive subsets, and even more of deciding when to stop splitting, are essential to the method. Often, the lack of general efficient stop splitting methods is evaded by means of alternative procedures, such as first building a very large tree, then pruning it; they generally rely on empirical justifications applicable to particular cases, but without guarantee of effectiveness in other application domains. To remove this difficulty, a stop splitting criterion was developed on the basis of a general statistical hypothesis test (Wehenkel *et al.*, 1989a, b). It was designed independently of any specific application, then applied to power system transient stability.

Transient stability in general is concerned with the system ability to withstand severe contingencies. A possible measure of this is the *critical clearing time* (CCT), i.e. the maximum time that a contingency may remain without causing the irrevocable loss of machines' synchronism. To compute CCTs one may either use time-domain methods, which solve numerically the nonlinear differential equations describing the system motion, or direct methods which rely on the Liapunov criterion (Bergen, 1986; Ribbens-Pavella and Evans, 1985). Observe that the CCT is not appropriate enough for assessing transient stability, in that it carries partly useless and partly incomplete information; indeed, it merely provides a rather crude "yes-or-no" answer, whereas what really matters in practice is to assess "stability margins" and, if necessary, to suggest remedial actions (EPRI Project, 1987). Within the tree methodology, however, the CCT provides a handy means of classifying the states of a LS.

* Received 13 June, 1989; revised 10 January, 1990; received in final form 25 January, 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor K.E. Arzen under the direction of Editor A. P. Sage.

† Department of Electrical Engineering, University of Liège, Institute Montefiore, B28, B-4000 Liège, Belgium.

‡ Research assistant F.N.R.S.

§ Author to whom all correspondence should be addressed.

The application of the tree methodology to transient stability was initially proposed in Wehenkel *et al.* (1986, 1987), then developed in Wehenkel *et al.* (1989a, b). The leading idea is to compress and organize the information about transient stability in the form of decision trees with the twofold objective: to classify new, unseen states, and to uncover the salient parameters driving the transient stability phenomena. Note that the foreseen advantages go beyond this objective: indeed, on-line means to compute stability margins and to infer control strategies were also suggested (Wehenkel *et al.*, 1988; Wehenkel, 1988). The first results obtained in a few practical examples were quite promising. The constructed trees exhibited nice features, notably with respect to accuracy and complexity. However, to make this method fully reliable and effective, a systematic exploration was necessary. This is a main purpose of the present paper.

Particular attention is paid to learning sets, "adequate" for building efficient decision trees; the way of generating their states in simulations and in real-life situations is discussed, and the influence of their number on the tree features is examined.

Another key issue concerns the splitting criterion. In Wehenkel *et al.* (1986, 1987), this consisted of a test applied to various "candidate attributes", chosen among static parameters of the power system, *a priori* likely to drive transient stability. Their influence on the tree structure is investigated below.

The cornerstone of the method is probably the stop splitting criterion (Wehenkel *et al.*, 1989a, b). Applied to transient stability, this criterion yielded simple, quite accurate trees in the few cases considered so far. In this paper, we determine its optimal parameters, and explore the accuracy and the complexity of the resulting trees.

The number of stability classes to be considered in the automatic building of trees is another question worth considering. Complexity, accuracy, misclassification rate and severity of misclassification errors are all interrelated features. Our purpose is to assess them, and to suggest possible practical applications, not necessarily to decide on a solution.

To conduct the above investigations we need appropriate tools for evaluating trees' accuracy. One of our first concerns will be to consider and compare *a priori* interesting evaluation criteria.

The ultimate goal for building trees is to use them. This topic is a whole research of its own. In this paper, we will merely indicate possible types of applications.

2. FUNDAMENTALS OF THE METHOD

2.1. General framework and basic notation

The automatic building of decision trees by the proposed inductive inference method implies the existence of a learning set, i.e. of a number, say N , of preclassified states. Without loss of generality, we will assume that each state is characterized by a certain number, say n , of ordered numerical *attributes* (the same number for each state), and that the N states are classified into two classes only $\{+, -\}$. The generalization to more than two-class classifications will be considered in Section 2.10; the extension to categorical attributes in addition to the ordered ones would be quite straightforward.

In the sequel, a learning set (LS) will be defined by:

$$LS \triangleq \{(\mathbf{v}_1, c_1), (\mathbf{v}_2, c_2), \dots, (\mathbf{v}_N, c_N)\} \quad (1)$$

where the components of vector \mathbf{v}_k

$$\mathbf{v}_k = (v_{1k}, v_{2k}, \dots, v_{nk})^T \quad (2)$$

represent the attribute values of the state s_k , which is characterized by its n attributes:

$$s_k = [a_1 = v_{1k}] \cap [a_2 = v_{2k}] \cap \dots \cap [a_n = v_{nk}], \quad (3)$$

and where

$$c_k \in \{+, -\}. \quad (4)$$

Note that the preclassified learning set is considered to be a statistical sample of size N , drawn from the population of possible states.

The test set (TS), is defined as a similar, but independent sample of size M :

$$TS \triangleq \{(\mathbf{v}_{N+1}, c_{N+1}), (\mathbf{v}_{N+2}, c_{N+2}), \dots, (\mathbf{v}_{N+M}, c_{N+M})\}. \quad (5)$$

It will be used for the purpose of evaluating the performance of a decision tree with respect to unseen states (see Section 2.8).

2.2. Decision trees (DTs)

A DT is a tree structured upside down. It is composed of *test* and *terminal* nodes, starting at the *top node* (or *root*) and progressing down to the terminal ones. Each test node is associated with a *test on the attribute values of the states*, to each possible outcome of which corresponds a successor node. The terminal nodes carry the information required to classify the states. Such a DT is portrayed in Fig. 1, built for an example treated in Section 4 in the context of transient stability. Note that in this case, where only numerical attributes are considered, the tests are dichotomic, and the resulting DT is binary: each test node is split into *two* successors.

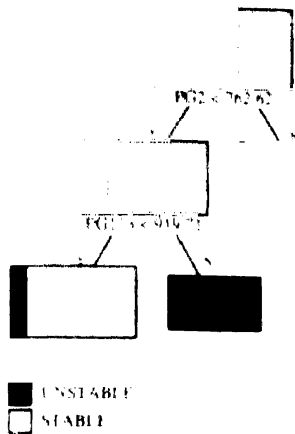


FIG. 1. A biclass tree for contingency #2. $N = 500$, $\alpha = 0.0001$

A convenient way to define a DT is to describe the way it is used for classifying a state of *a priori* unknown classification on the basis of its known attributes. This classification is achieved by applying sequentially the test at the test nodes, beginning at the root of the tree, and systematically passing the state to the successor appropriate to the outcome of the test, until a terminal node is finally reached; the state is classified accordingly.

The above description provides an interesting, geometric interpretation of a tree procedure (Breiman *et al.*, 1984): it recursively partitions the attribute space into hyperboxes, such that the population within each box becomes more and more class homogeneous. This in turn allows one to view the classification of a tree as the partitioning of this hyperspace into two regions, corresponding to the two classes. Each class is composed of the union of the elementary boxes corresponding to its terminal nodes. Figure 2 illustrates the geometric representation of the tree of Fig. 1. (Actually, according to this tree structure, the states belonging to the region labelled "stable" have a small chance (13%) to be unstable.)

2.3. Automatic construction of DTs

For a given LS, our purpose is to build a near optimal DT, in the sense that it realizes a good tradeoff between complexity and accuracy, i.e. between total number of nodes and classification

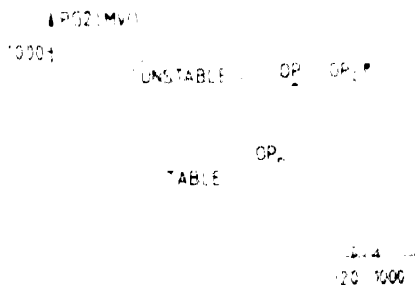


FIG. 2. Geometric representation of the tree of Fig. 1

ability. Usual ways of estimating accuracy are specified in Section 2.8; anticipating, observe that, intuitively, assessing trees' ability to correctly classify states should rely on test states rather than on states belonging to LS itself.

The ultimate objective of the building procedure is to: (i) select the relevant attributes among the candidate ones (generally the number of retained attributes is much lower than n , the total number of candidate attributes); (ii) build the decision tree on the basis of these relevant attributes. This building procedure is described below (Wehenkel *et al.*, 1989a).

- Starting at the root of the tree, with the list of candidate attributes and with the whole LS, the learning states are analysed in order to select a test which allows a maximum increase in purity or, equivalently, which provides a maximum amount of information about their classification. The selection proceeds in two steps.

- for each attribute, say a_i , it finds the optimal test on its values, by scanning the values of this candidate attribute, for the different learning states, in our case of ordered numerical attributes, this step provides an optimal threshold value v_{i*} and defines the test

$$a_i \leq v_{i*} \quad (6)$$

- among the different candidate attributes, it chooses the best one, a_{*} , along with its optimal value, v_{*} , to split the node.

In short, step (ii) defines the optimal attribute, step (i) its optimal threshold value

- The selected test is applied to the learning set of the node and splits it into two subsets, corresponding to the two successors of the node. Starting with the root of the tree and the entire LS, the two subsets

$$LS_1 \triangleq \{v_k \in LS | a_{*} \leq v_{*}\}$$

$$LS_2 \triangleq \{v_k \in LS | a_{*} > v_{*}\},$$

correspond to the two successors of the root.

- The successors are labeled terminal or not on the basis of the *stop splitting criterion* described below.
- For the nonterminal nodes, the overall procedure is called recursively in order to build the corresponding subtrees.
- For the terminal nodes, the class probabilities p_+ and p_- are estimated on the basis of the corresponding subset of learning states there stored, and the class label of the majority class is attached.

Obviously, the crux of the entire construction of a DT lies in the selection of the splits and the

of a DT lies in the selection of the splits and the decision whether to declare a node terminal or to continue splitting. These two questions are examined below, after the introduction of some necessary notions of (im)purity and information, drawn from the information theory.

2.4. Mathematical formulation

2.4.1. Definitions. Denoting by:

S the subset of all possible states, corresponding to some node of a DT, i.e. directed to that node by the classification procedure;

$p(+ | S)$ (resp. $p(- | S)$) the probability of a random state drawn from S to belong to the class $+$ (resp. $-$);

T a dichotomic (candidate) test [e.g. see (6)] on some attribute's values of the states;

S_v (resp. S_n) the subsets of S of states yielding the answer "YES" (resp. "NO") to the test T ; these sets would correspond to the successors of the node split on the basis of T ;

$p(S_v | S)$ (resp. $p(S_n | S)$) the probabilities of the outcomes "YES" and "NO" in S ; one defines the following measures.

The *prior* "classification entropy" of S

$$H_c(S) \triangleq -[p(+ | S) \log_2(p(+ | S)) + p(- | S) \log_2(p(- | S))]. \quad (7)$$

$H_c(S)$ is a measure of the impurity or the uncertainty of the classification of a state of S ; $H_c(S) = 0$ corresponds to a perfectly pure ($p(+ | S) = 1$ or 0) subset, whereas $H_c(S) = 1$ corresponds to $p(+ | S) = p(- | S) = 1/2$ i.e. maximal uncertainty.

The entropy of S with respect to the test T

$$H_T(S) \triangleq -[p(S_v | S) \log_2(p(S_v | S)) + p(S_n | S) \log_2(p(S_n | S))]. \quad (8)$$

$H_T(S)$ is a measure of the uncertainty of the outcome of T in S ; with respect to the outcome of T , it has similar properties to H_c with respect to the classification.

The mean *posterior* "classification entropy" of S given the outcome of T

$$H_{c|T}(S) \triangleq p(S_v | S)H_c(S_v) + p(S_n | S)H_c(S_n). \quad (9)$$

$H_{c|T}(S)$ is a measure of the residual impurity if S is split into S_v and S_n according to the outcomes of test T .

The "information" provided by T on the classification of S

$$I_c^T(S) \triangleq H_c(S) - H_{c|T}(S); \quad (10)$$

$I_c^T(S)$ is a measure of the ability of T to produce pure successors.

The normalized information gain of T

$$C_c^T(S) \triangleq \frac{2I_c^T(S)}{H_c(S) + H_T(S)}. \quad (11)$$

Remark. The measures defined in (7)–(10) are expressed in bits whereas that in (11) is dimensionless.

2.4.2. Interpretation. It is possible to give a qualitative interpretation of the relation between $C_c^T(S)$ and the classification ability of the test T . To this, first observe that the following inequalities can be easily shown:

$$0 \leq H_c(S), H_T(S), H_{c|T}(S), I_c^T(S), C_c^T(S) \leq 1 \quad (12)$$

$$H_{c|T}(S), I_c^T(S) \leq H_c(S). \quad (13)$$

Then consider the two following extreme cases:

1. The outcome {"YES", "NO"} of T and the classification $\{+, -\}$ are statistically independent, i.e. the test T provides no information at all about the classification. In this case

$$p(+ | S_v) = p(+ | S_n) = p(+ | S) \quad \text{and} \\ p(- | S_v) = p(- | S_n) = p(- | S). \quad (14)$$

Thus, $H_c(S_v) = H_c(S_n) = H_c(S)$ and according to (9), $H_{c|T}(S) = H_c(S)$. In addition, by virtue of (10) and (11),

$$I_c^T(S) = C_c^T(S) = 0.$$

2. The outcome of T provides pure classified subsets S_v and S_n . Then necessarily $H_c(S_v) = H_c(S_n) = 0$. According to (9), this implies that $H_{c|T}(S) = 0$ and $I_c^T(S) = H_c(S)$. Moreover, in this case the probability $p(S_v | S)$ (resp. $p(S_n | S)$) will be equal to either $p(+ | S)$ (resp. $p(- | S)$) or $p(- | S)$ (resp. $p(+ | S)$), and thus $H_T(S) = H_c(S)$ and

$$C_c^T(S) = 1.$$

From the above it follows that the higher the value of $C_c^T(S)$, the more interesting the test T for splitting the node corresponding to S . This is exploited in Section 2.5 to define the optimal splitting rule, used in the building of DTs.

2.4.3. Estimation on the basis of the sample of learning states. The information and entropy measures defined in (7)–(11) cannot be computed directly since the probabilities involved are generally unknown. Therefore, we use the learning set as a statistical sample and estimate the probabilities by their empirical values computed in the LS. More precisely, the set S , corresponding to some node of the DT, is replaced by the *subset* of learning states (i.e. $S \cap LS$) corresponding to this node. The computation of $p(+ | S)$, $p(- | S)$, $p(S_v | S)$ and $p(S_n | S)$ of this subset is then straightforward, since the classification and attribute values of the learning states are known.

It should be noted that, although the estimates of $p(+ | S)$, $p(- | S)$, $p(S_v | S)$ and $p(S_n | S)$ are

generally unbiased, their substitution in (7)–(11) provides rather optimistically biased information measures, thus overestimating the actual “goodness” of the test T . Fortunately, the amount of bias is inversely proportional to the sample size (the number l of states in $S \cap LS$) and the measures may still be used for comparison purposes, e.g. in order to select an “optimal” test for splitting the node.

For example, it can be shown (Kvålseth, 1987) that for an actual value of $I_1^l(S) = 0$ (no correlation), its estimate

$$\hat{I} = I_1^l(S \cap LS) \quad (15)$$

has a χ -square like distribution, the expectation (or bias) of which is positive and inversely proportional to the number l of states in $S \cap LS$. This property is exploited below, in the stop splitting criterion.

In the sequel, to emphasize the difference between the purity measures and their estimates, we will call the latter “apparent” as opposed to the “real” unknown values.

2.5. On the optimal splitting criterion

Obviously, the splits at the test nodes should be selected so as to avoid to the extent possible the appearing of deadends (i.e. of impure terminal nodes, see the definition given below in Section 2.6), and to obtain the desired degree of accuracy. This is done in a more or less heuristic (not necessarily optimal) fashion: the best test (defined by the optimal attribute together with its optimal threshold value, see Section 2.3), is considered to be the one which separates at most the states of the two classes in the local learning subset. This strategy amounts to selecting the split which yields the purest direct successors, or maximizes the apparent normalized information gain $C_1^l(S \cap LS)$ defined in (11). In that sense, it may be considered to be *locally*, rather than *globally* optimal.

2.6. Stop splitting method

Many methods were proposed. For example, ID3 (Quinlan, 1986) stops splitting at a node only if the corresponding learning subset is completely included in one of the classes of the goal partition. Unluckily, in many situations, as shown by our experience, this strategy tends to build overly complex DTs, most of the terminal nodes of which contain only a very small and unrepresentative sample of learning states. They perform generally badly with respect to unseen situations and are unable to indicate in a reliable way the inherent relationship between the attributes and the goal classification. To circumvent this difficulty we propose a more

conservative criterion, which stops splitting a node if one of the following two conditions is met:

- 1 The local subset of learning states is sufficiently class pure. Such a terminal node will be called a *leaf* in the sequel. The degree of class purity required for leaves is a parameter of the algorithm and fixes the amount of detail we want the DT to express.

Note Actually, the degree of residual “impurity” may be specified by H_m , the maximal residual entropy [see (7)]. The entropy of the learning subset relative to a node, is inversely proportional to its purity. Thus, in terms of entropy, a node will be a leaf only if its entropy is lower than H_m . In the practical investigations reported in this paper, a constant value of $H_m = 0.1$ bits was used. This very low value amounts to building DTs as detailed as possible. Higher values could be of interest if one wanted to build simpler “first guess” DTs, for data exploration purposes.

- 2 There is no possibility to enhance the DT accuracy in a statistically significant way by splitting this node further. In other words, given the optimal dichotomic split for this node, there is not enough evidence in the local learning subset, that this split would actually improve the real performance of the DT. Such a terminal node is called a *deadend* in the sequel. This second criterion prevents the building of unnecessarily complex DTs. It is formulated as a statistical hypothesis test:

Given the local subset of learning states and the optimal split, can we accept the hypothesis that the apparent increase in accuracy, as measured in this subset, provided by the split, is a purely random effect?*

In quantitative terms, the statistic $\hat{A} \hat{\Delta} c/l$ (where l is the size of the local learning subset, c a constant and \hat{A} the apparent increase in purity provided by the optimal split) is distributed according to a χ -square law (with one degree of freedom) under the hypothesis of no *real* increase in purity. Hence, if we fix the α -risk of not detecting these situations to a given value, testing the value of \hat{A} against the threshold \hat{A}_c , such that $\text{Prob}(\hat{A} \geq \hat{A}_c) = \alpha$ allows one to detect the cases where the apparent increase in accuracy is a random effect, with a probability of $1 - \alpha$. Figure 3 sketches such χ -square probability density functions.

* Apparent as opposed to real increase in accuracy, means the increase measured in the LS as opposed to the increase measured for unseen states.

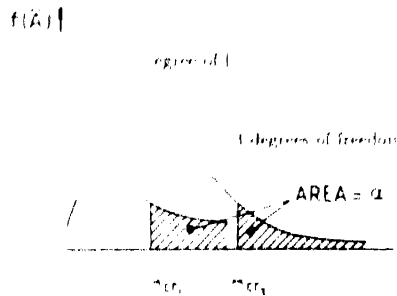


FIG. 4. χ -square probability density functions of \hat{A}

Thus, the α -risk of the hypothesis test fixes the amount of evidence we require at each node in order to split it; the answer depends on the value of α , the size of the local learning set, and the amount of apparent accuracy improvement provided by the test. The question of “how much evidence should be required to allow the splitting of a node”, is related to the degree of representativity we impute to the learning set and the risk that this degree is overestimated. Therefore, the degree of cautiousness of the procedure is fixed by the user via the selection of the value of α . This value ranges from 1 (the criterion has no effect on the splitting procedure anymore, and the tree grows according to the above condition 1) to zero (no growth is allowed, the tree reduces to its root).

The value of α has indeed drastic effects on the resulting trees characteristics as illustrated by the following example (borrowed from Section 4, below). Four DTs were grown for α assuming the values 1.0, 0.1, 0.01 and 0.0001; they were built on the basis of a LS composed of 500 states and evaluated on the basis of an independent test set of 1500 other states. Table 1 reports their complexity and accuracy, i.e. the number of their nodes and the percentage of correctly classified test states. These quite impressive figures illustrate that the larger trees are less accurate than the smaller ones and on the other hand that the appropriate selection of α allows the building of small, yet better DTs (see also the discussion of Section 2.9 and the simulation results of Section 4.3.2).

2.7. Comparison with the ID3 method

The proposed inductive inference method originates from the ID3 algorithm, from which, however, it departs in some essential respects.

Below, we collect the main differences between the two methods.

Application domain. ID3 was initially intended to handle mainly *symbolic* and *deterministic* learning problems characterized by very large (almost complete) learning sets composed of objects described by discrete (or qualitative) attributes only. Thus, it was essentially designed to handle large sets of data, in order to *compress* rather than extrapolate their information (Quinlan, 1984). On the contrary, the proposed method is especially tuned to handle mainly *numeric* and *nondeterministic* problems, where the learning set has to be *generalized* in an appropriate fashion. The method is general enough to handle at the same time numeric and qualitative attributes and can be tuned to the degree of “representativeness” of the learning set, via the appropriate choice of the threshold α (see Section 4.3 below).

Stop splitting criterion. ID3’s stop splitting criterion amounts to building DTs which classify the learning set as correctly as possible, which is indeed the best approach if the latter is almost complete. The proposed method, on the other hand, stops splitting earlier, as soon as the statistical hypothesis test allows to conclude that no *significant* improvement of the tree’s accuracy would be achieved by developing the node anymore. This is quite different, and enables the method to take the best advantage of learning sets which are only partly representative of the possible objects. This is further discussed in Section 2.9 below.

Optimal splitting criterion. ID3 considers the best test to be the one providing the largest apparent information gain \hat{I}'_t . It has been found that this measure is biased in the favour of those having the largest number of successors, especially in the context of randomness (Quinlan, 1986). Note that this is supported by the fact that the number of degrees of freedom of the χ -square distribution (and thus its bias) increases linearly with the number of successors of a split. The *normalized* correlation measure \hat{C}^T_t evaluates the tests more objectively since the number of successors is compensated by the term \hat{H}_T in its denominator.

Strategy. ID3’s strategy, embodied by its optimal and stop splitting criteria, amounts to building trees which maximize the *apparent reliability*, regardless of their complexity. On the other hand, the strategy of the proposed method is to build DTs maximizing a measure of *quality* which realizes a compromise between *apparent reliability* and *complexity*. The latter is more effective in the context of incomplete and random data (Wehenkel, 1990).

TABLE 1

	1.00	0.1	0.01	0.0001
Complexity	63	55	5	5
Accuracy	88.1	88.5	91.2	91.2

2.8. Evaluation criteria

The criteria described below (e.g. see Toussaint, 1974) allow assessing the accuracy of a DT in general, or equivalently, its misclassification rate. Observe that these are only estimates, since in real world problems it is seldom possible to scan all the possible states, even in deterministic type problems.

Resubstitution estimate. R^b obtained by considering all the states belonging to the LS, dropping them through the DT, and computing the ratio of misclassified over the total number (N) of states:

Test sample estimate. R^b obtained by considering all the states belonging to the TS, dropping them through the DT, and computing the ratio of misclassified over the total number (M) of states;

Cross-validation estimate. R^b obtained by: (i) dividing the total number of learning states into, say V , equally sized subsets, using one of them as the test set and the union of the remaining $V - 1$ as a learning set; (ii) building a DT based on this learning set and assessing its accuracy on the basis of the test set; (iii) repeating this procedure V times by considering successively each of the V subsets as test set; (iv) taking the average of the V individual estimates as the final accuracy. The validity of the procedure implies that V is quite large (> 10), so that the DTs constructed on the basis of the $(V - 1)$ subsets are (almost) identical with the initial DT, constructed on the basis of the total number of states (V subsets). For $V = N$ this is the so-called "leave-one-out" estimate.

Remark. The above defined measures can be considered as more or less accurate estimates of the true probability of misclassifying a new state. They have the following characteristics:

R^b is easy to compute and requires no additional samples but is generally optimistically biased, underestimating the real probability of error. Intuitively, for fixed N , the larger the DT, the higher the correlation between the DT and the LS, and the more biased R^b .

R^b is the most reliable and unbiased estimate and is easy to compute. Moreover, if M is sufficiently large its variance is small. Its main drawback is that it requires additional states in sufficiently large number, which cannot be used in the learning set. Another advantage of R^b is that one can easily estimate its variance and therefrom compute confidence intervals. In the sequel we will consider this estimate as the benchmark.

R^b combines advantages of the two preceding measures, since it is generally less biased than R^b and does not require additional states. On the other hand, its variance can be very large, and certainly depends strongly on the characteristics of the problem at hand. In practical situations, when the number of samples is limited, it is an interesting alternative to R^b . Notice also that the computational burden required for R^b can be very heavy, since it needs the building of V additional DTs.

The above considerations are illustrated in Section 4, on a real world example.

2.9. On the right size of trees

The stop splitting method introduces the statistical threshold parameter α . Varying its value allows to modify the size of a tree. But how to choose the right tree size? The choice should be guided by the observation that too large a DT will yield a higher misclassification rate and hide relevant relationships, whereas too small a tree will not make use of some of the information available in the LS. Indeed, the terminal nodes of a too small DT have not been expanded enough and this prevents from getting the purer subsets and the corresponding insight about the role that the attributes would have played in this expansion, a too large DT, on the other hand, results from the splitting of statistically unrepresentative subsets; therefore, it is likely to cause an increase in the misclassification rate when classifying states not belonging to the LS. Stated otherwise, a tradeoff appears between the two following sources of misclassification: *bias* (overlooking significant information in the LS) and *variance* (badly interpreting the randomness in the LS); too large a tree will suffer from variance whereas too small a tree will present bias.

2.10. Building multiclass DTs

The above description of the inductive inference method made in the case of two classes $\{+, -\}$, is general enough to handle (at least in principle) an arbitrary number, say m of classes, provided that m remains negligible with respect to the size N of the LS.

Indeed, on one hand the information theoretic purity measures defined in (7)–(11) may be generalized to m classes; on the other hand the statistical hypothesis test remains applicable, provided that $m - 1$ degrees of freedom are used (instead of 1 in the two-class (or biclass) case) for the χ -square law.

Under these conditions, the method will build

DTs classifying directly the states into one of the m specified classes.

Another indirect possibility would be to build $m - 1$ biclass DTs and to combine them in order to obtain the m -class classification.

In the investigations of Section 4 we use the first, direct approach. The obtained DTs are simpler and easier to interpret. Moreover, preliminary investigations indicate that they are at least as good, sometimes better, from the accuracy viewpoint, than the indirect multi-biclass trees.

3. DECISION TREE BASED TRANSIENT STABILITY APPROACH

Two main conjectures underlie the application of the tree method to transient stability assessment. First, transient stability is strongly dependent upon the contingency type and location; hence the idea of building a tree per contingency. Second, transient stability is a quite localized phenomenon, and is driven by a few number of the system parameters; hence the idea of proposing candidate attributes selected among the parameters of the system in its steady state condition.

These generally well-accepted conjectures, have also been verified in the few cases treated by the tree methodology (Wehenkel *et al.*, 1989a, b). The case study of the next section attempts to further validate them by means of exhaustive simulations. It also provides answers to questions raised by the overall *decision tree transient stability* (DTTS) method, whose principle is recalled below.

3.1. Principle of the DTTS approach and related questions

This may be formulated as follows: for each preassigned contingency, build up off-line a DT on the basis of a learning set and of candidate attributes. This tree is then used on-line to classify new, unseen states in as many classes as desired; for example, in a biclass tree, a given state would be classified as either stable or unstable, whereas in a three-class tree, the same state would be declared stable, fairly stable, or unstable.

Below, we identify questions relevant to this definition and suggest answers, often anticipating the results of Section 4.

3.1.1. Questions about the LS.

- (i) How to obtain the learning states;
- (ii) How to classify them;
- (iii) How many states should be used for the efficient construction of a DT;
- (iv) Whether the "right" size of LS should be

dependent upon the size of the power system.

Tentative answers.

- (i) Either by considering plausible scenarios and running a load flow program to get the corresponding operating states, or by using past records of the system real life;
- (ii) According to their CCT values, computed via a time-domain or a direct method as appropriate;
- (iii) This question is explored in the next section;
- (iv) The answer to this question is postponed until Section 5.

3.1.2. Questions about the list of candidate attributes.

- (i) What kind of candidate attributes to select for test (6);
- (ii) How many;
- (iii) What would happen if the actually most relevant attribute were for any reason masked, i.e. discarded from the list;
- (iv) What if additional relevant attributes are further considered.

Tentative answers.

- (i) Parameters of the system in its steady state, pre-contingency, operation;
- (ii) When no prior knowledge of the system can guide the selection, it is advisable to consider as many as possible candidates of the type just suggested, confined in a relatively restricted area surrounding the contingency location; note that the increase in computing cost is insignificant (see below);
- (iii) It would generally lead to somewhat more complex, yet quite accurate DTs, provided that other relevant attributes are still present in the list;
- (iv) See results of Section 4.3.4.

3.1.3. Questions relating to the number of classification patterns.

- (i) Which are the main differences between trees of small and large number of classes;
- (ii) For a given misclassification rate, which type of trees provides narrower misclassification error.

Tentative answers

- (i) One can reasonably conjecture that the smaller the number of classes in a tree, the less complex and more accurate this tree;
- (ii) Provided that the inductive inference method is correctly developed, it is normal to think that the larger the number of classes in a tree, the less severe the misclassification

errors; indeed, in a well designed tree, misclassification errors will result in adjacent classes; it is therefore less severe to declare fairly stable a state which actually is stable (three-class tree), than to declare it unstable instead of stable (two-class tree) (see Section 4.3.5).

3.1.4. Questions relating to computational aspects.

- (i) Which is the most expensive task of the DTTS approach;
- (ii) How does the number of candidate attributes affect the computing time of a DT;
- (iii) How does the number N of learning states affect the computing time of a DT;
- (iv) How "expensive" is the storage of DTs;
- (v) How "expensive" is their use

Tentative answers.

- (i) The generation of a LS is undoubtedly the most demanding task; and the more refined the system modelling, the more expensive the task;
- (ii) According to the splitting criterion procedure described in Section 2.3, the time required by a DT building varies linearly with the number of candidate attributes;
- (iii) In terms of the size of the learning set, the computing time is upper-bounded by—and generally much lower than— $N \log N$, i.e. the time required to sort the LS according to the values of the n candidate attributes;
- (iv) The storage of a DT is generally extremely inexpensive; indeed, it is proportional to the number of its nodes, which is found to be quite small (see next section); to fix ideas, a compiled LISP version of a DT-classifier composed of 25 nodes (which can be considered as an upper bound for the DTTS method) takes about 600 machine instructions, meaning that in the context of modern computer architectures thousands of DTs can be stored simultaneously into main memory;
- (v) The mean classification time of a state by means of a DT is almost negligible (about 0.6 ms for a DT comprising 25 nodes)

3.2. Possible types of uses of the DTs

Various uses can be inferred by following the pattern of the general DT methodology (e.g. see Section 2.2); many others are specific to the DTTS approach.

The use which immediately comes to mind is the classification of an operating state of *a priori* unknown degree of stability with respect to a

given contingency: considering the appropriate DT, and applying to this state the test (6), one successively directs it through the various nodes of the tree, until reaching a terminal node, the state is classified accordingly. One could object that the information thus obtained is less refined than the CCTs used to classify the states of the LS; this however is not surprising, since the precise CCT values of the learning states are not fully exploited during the tree building. Wehenkel *et al.* (1988) and Wehenkel (1988) proposed a means to approximately estimate the CCT of the state one seeks to classify, by using the notion of "distance to a class", whose definition takes into account the CCTs. This distance, inferred from the geometric interpretation of a DT outlined in Section 2.2, is schematically illustrated in Fig. 2. One can see that the *operating* state (or *point*, OP) labelled OP_n is stable, whereas the state OP_1 is unstable and OP_2 is very unstable. Moreover, it shows that to (reinforce stability, one should suitably direct the OP in the attribute space, i.e. suitably modify the values of its relevant attributes.

The above short description suggests that the DTTS approach can provide three types of transient stability assessment: analysis, directly linked to the classification of an OP, sensitivity analysis, linked to the "distance of an OP to a class", and control, linked to the way one can act to modify such a distance. Moreover, since the involved computations are likely to be extremely fast, several real-time strategies may be conceived. For example, one may proceed in a way similar to the standard "contingency evaluation" of steady state security assessment, viz.: draw up a contingency list, and build off-line the corresponding DTs, then, in real-time, scan the list, focus on those contingencies which are likely to create problems, and if necessary propose corrective actions. Of course, care should be taken so as to avoid incompatible actions. But for the time being this question is beyond our scope.

Finally, observe that the DTs provide also a clear, straightforward insight into the complex mechanism of transient stability. This is not the least interesting contribution of the DTTS approach

4. CASE STUDY

4.1. General description

The investigations conducted in this section aim to answer the questions previously raised, to assess the salient features of the DTTS approach and to test the conjectures underlying it.

The power system we have chosen is small enough to avoid unnecessarily bulky computa-

tions, but large enough to draw realistic conclusions. It is composed of 31 machines, 128 buses and 253 lines; its total generation power in the base case amounts to 39,000 MW. This system is described in Lee, (1972), and sketched in Fig. 4. (The rationale of its decomposition is discussed in Section 4.2.1.)

In this first set of large-scale simulations we adopted the standard simplified system modelling, i.e. constant electromotive force behind transient reactance for each machine and constant impedance for each load.

To assess the DTTS method as objectively as possible, we have generated 2000 operating points (OPs). Two sets of simulations were considered, corresponding to very severe contingencies, consisting of three-phase short-circuits (3 ϕ SCs) applied at generator buses (one at a time). The first set concerns thorough investigations carried out with three such stability scenarios, of 3 ϕ SCs applied at buses #2, 21, and 49, arbitrarily chosen.* This implied the computation of 6000 CCT values. Part of the OPs along with their classification have been used in the LS, the other part composing the TS. A large number of DTs have been built for the above three contingencies, for various sizes of LSs, varying values of the threshold α , and three different classification patterns, namely two-, three-, and four-class classifications. In the two-class case, one distinguishes stable and unstable OPs, depending upon whether their CCT is above or below a certain threshold CCT value; in the three-class case one distinguishes stable, fairly stable and unstable OPs (two threshold CCT values are used); in the four-class case one distinguishes very stable, stable, unstable and very unstable OPs (three threshold values). The threshold CCT values are *a priori* chosen rather arbitrarily, but so as to avoid a too important imbalance between the populations of OPs belonging to the various classes.

The second set of simulations is described in Section 4.4.

4.2. Constitution of a data base

The data base was randomly generated on the basis of plausible scenarios, corresponding to various topologies, load levels, load distributions and generation dispatches. Hereafter we describe the way used to generate them, to analyse them from the transient stability point of view, and to build the attribute files.

4.2.1. *Random generation of OPs.* To generate these various states, we grouped the nodes

of the power system into five internal zones, and one external zone composed of the boundary nodes of the system and its external equivalents (Fig. 4). The internal zones were defined empirically, on the basis of the "electrical distances" (number and length of lines) connecting their nodes. The OPs composing the data base were generated randomly according to the following independent steps.

1. *Topology:* it is selected by considering base cases (with probability 0.5) and single outage of a generator, a load or a shunt reactor (each with probability 0.08), a single line (with probability 0.16), two lines (with probability 0.1). The outaged element is selected randomly among all the elements of the power system.
2. *Active load level:* the total load level is defined according to a Gaussian distribution (with $\mu = 32,000$ MW and $\sigma = 9000$ MW).
3. *Distribution of the active load:* the total load is first distributed among the six zones according to the random selection of participation factors (see below), then among the load buses of each zone homothetically with respect to their base case values. The reactive load of each bus is adjusted according to the local base case power factor. This results in a very strong correlation of the loads of a same zone, and a quite weak one among loads of different zones.
4. *Distribution of the active generation:* in a similar fashion, the total generation is first distributed among the zones according to randomly selected participation factors, then among the generators of each zone according to a second selection of participation factors. Thus, neighbour generators are less correlated than neighbour loads. The reactive generations are obtained by a load flow calculation. To avoid overloads, the final active generation of each generator is constrained to 90% of its nominal power.

Note. In order to avoid overloading or underloading the slack generator, the total generation is defined as the total active load plus a polynomial approximation of the active network losses of form:

Network Losses (MW)

$$\approx 1336 - 255P_1 + 19P_1^2 - 0.7P_1^3 + 0.012P_1^4 - 0.00007P_1^5$$

where P_1 denotes the total active load of the OP, in GW, and where the coefficients were determined by a least squares estimation on the basis of 11 OPs of different load levels for which the network losses were computed by a load flow program.

* In the sequel, a contingency will often be specified by merely the number of the generator bus at which it is supposed to apply; e.g. contingency (or fault) #2, #21, #49.

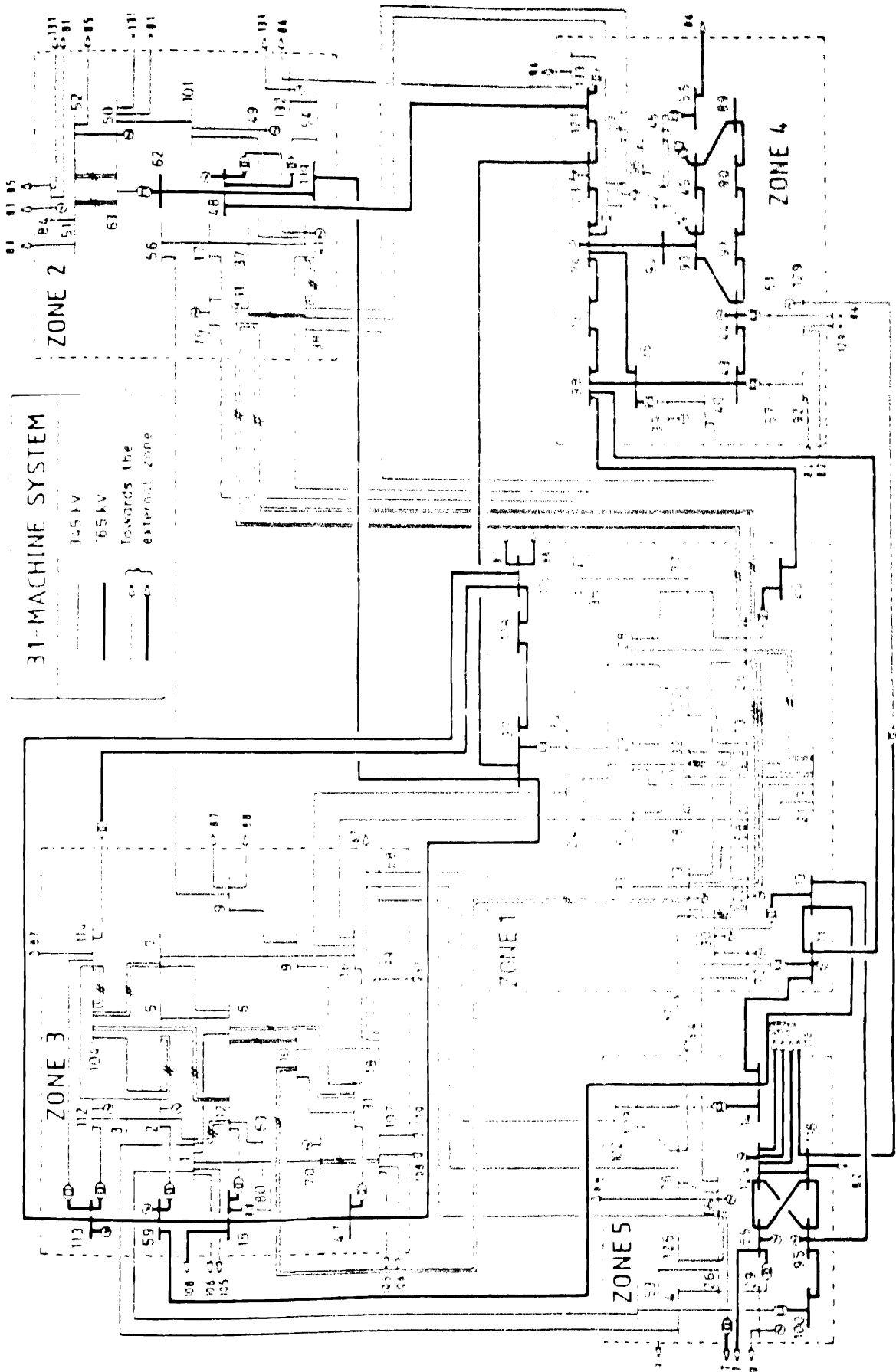


FIG. 4. One-line diagram of the 31-machine test system and its decomposition

5. Load flow calculation: to check the feasibility of an operating point and compute its state vector, it is fed into the load flow program, and accepted if the latter converges properly. (About 90% of the states were accepted.) (Remember that the power system has 128 buses and hence each state vector has $2 \times 128 - 1 = 255$ components.)

Participation factors. To distribute a given amount of active (generation or load) power (say P_{tot}) on a number, say E , of "elements" (zones, loads or generators) we use the following procedure.

Let P_i^N , ($i = 1, \dots, E$) be the "nominal" power of the elements (for zonal loads we use the base case total load of the zone, for generators their nominal power and for zonal generations the sum of the nominal generations of their generators), and λ_i , ($i = 1, \dots, E$), positive coefficients selected randomly, according to the procedure described below. The individual participation of each element is defined by:

$$P_i \triangleq \frac{\lambda_i P_i^N}{\sum_{i=1}^E \lambda_i P_i^N} P_{tot}. \quad (16)$$

Thus, the λ_i coefficients act as a distortion with respect to the homothetic nominal power distribution. They are selected randomly in the following way:

One third of the cases are defined by $\lambda_i = 1$, ($i = 1, \dots, E$), i.e. no distortion with respect to the nominal distribution;

The rest of the cases are defined by $\lambda_i = 2$, for a randomly selected value of j , and $\lambda_i = 1$, ($i = 1, \dots, j-1, j+1, \dots, E$), i.e. a higher participation of the j th element.

4.2.2. Transient stability (pre)analysis and attribute calculation. To use the data base for investigating the DTTS approach, we carried out the following preliminary calculations:

1. Approximate calculation of the CCT of the 2000 OPs, for a 3 ϕ SC at each one of the 31 generator buses, using the extended equal area criterion (EEAC) (Xue *et al.*, 1988). This gave us good information about the relative severity of these contingencies in relation to the OPs represented in the data base, and allowed us to select three "interesting" ones for our investigations.

2. Precise calculation of the CCTs of the OPs, for the three selected faults, using the step by step (SBS) method. To accelerate the iterative cut and try process delimiting the precise value of the CCT, the approximate values supplied by

the EEAC were used as an initial guess. Incidentally, these latter were found to be in a very good agreement with those computed by the SBS method (over the 6000 CCT values we found a negative bias of $-0.007s$ and a standard deviation of $0.032s$ of the CCTs provided by the EEAC as compared to those computed by the SBS method).

3. Generation of the files containing the attribute values for the 2000 OPs. About 270 different "primary" attributes have been computed, comprising zonal statistics on loads, generation and voltage, voltage magnitudes at all buses, voltage angles at important buses, active and reactive power of each generator, and topology information for each OP. Other attributes, such as line power flows, could be defined as simple algebraic combinations of the primary attributes and did not necessitate to be stored explicitly. The attribute files were constructed on the basis of the load flow data and state vectors.

Overall, the data base contains about 300 values per operating point, in addition to the input files required for the load flow and transient stability analysis programs.

4.3. Simulation results

Over 400 different DTs were built for various scenarios: three different contingencies, about 15 different classifications, distributed almost equally among the two-, three-, and four-class patterns, learning set sizes ranging from 100 to 1500 OPs, more than 100 different candidate attributes, α values ranging from 1.0 to 0.00005. The DTs were evaluated on the basis of independent test sets.

The resulting observations are organized and presented in six parts (Sections 4.3.1–4.3.6), although they are interdependent in many respects. The first part analyses general DT features, such as complexity and accuracy, with respect to various classification patterns, while fixing the other parameters (size of the LS, value of α , list of candidate attributes); this provides a good insight into the overall DTTS method. The second part focuses on the influence of α on the resulting DTs, and suggests how to select good α values in practical situations. The third part explores the way the size of the LS affects the DTs, while the fourth part discusses the influence of the candidate attributes on the DTs. The fifth part examines the influence of the number of classes on the misclassification rate and severity. Finally, the sixth part compares the different accuracy estimates defined in Section 2.8.

TABLE 2. TREE FEATURES AS RELATED TO THE NUMBER OF CLASSES. $\alpha = 0.0001$, $N = 500$, $M = 1500$

Gen bus #	Two classes			Three classes			Four classes		
	N_t	$R^b(\%)$	A	N_t	$R^b(\%)$	A	N_t	$R^b(\%)$	A
2	7	2.27	3	17	5.67	4	27	9.60	7
21	9	3.73	3	17	3.60	4	25	8.40	6
49	5	1.53	1	9	4.33	2	15	7.53	3

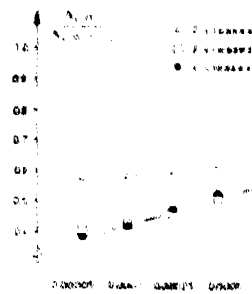
4.3.1. *General tree characteristics.* DTs corresponding to three different classifications patterns (of two, three and four classes) are built for each of the three different contingencies under the following conditions: the data base (2000 OPs) is divided in a LS composed of $N = 500$ OPs and a test set composed of the $M = 1500$ remaining OPs (this fairly large size provides a high precision to the R^b estimate); the value of α is fixed to 0.0001 (the justification of this choice will be found below); and the list of candidate attributes, the same for each DT, is composed of 81 static variables.

The results are summarized in Table 2. The first column specifies the faulted bus number of the contingency. The nine following columns specify for the indicated contingency and number of classes, the characteristics of the resulting DT: N_t , the total number of its nodes (which measures its complexity); R^b , its test sample estimate, representing the percentage of misclassified test states; A, the number of different retained attributes among the 81 candidates. This is repeated for the three contingencies, providing the features of nine DTs listed in the table.

The same set of investigations was repeated three more times, with three other learning sets of the same size (each of 500 states), in order to detect the variability of the DTs with the LS. The obtained results are very similar to the above, and induce the following conclusions:

- The complexity increases from the very simple two-class trees to the moderately complicated three- and four-class ones;
- Their accuracy is quite satisfactory especially in the two- and three-class cases; the four-class trees are less accurate but, as we discuss below, their errors are less harmful;
- The error rate varies only moderately with the fault location;
- The total number of retained attributes remains overall very small.

This latter aspect justifies the conjecture that transient stability is a localized phenomenon and highlights the ability of the method to select the relevant attributes. (A worth mentioning fact

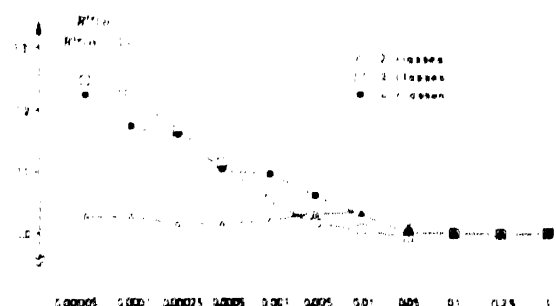
FIG. 5. Influence of α on the normalized number of nodes $N = 500$, $M = 1500$

which does not appear in the table, is that the selected attributes are essentially variables related to buses very close to the faulted one.)

4.3.2. *Impact of the parameter α .* For a given LS, attributes list and classification pattern, the principle of the stop splitting criterion suggests that the lower the value of α the smaller the resulting DT. Although this reduction in complexity is certainly interesting from many viewpoints, it could also cause the DTs to be less accurate, as indicated by our earlier discussions. Hence, the necessity of scrutinizing the effect of α on the complexity and accuracy of the DTs.

These observations were extensively investigated using different faults, numbers of classes, and learning sets. They yielded the following conclusions.

- The most important reduction of a DT's complexity is obtained as soon as α enters the range 0.001 to 0.0001, quantitatively, this effect is more marked in the three- and four-class cases.
- Although lower values of α still reduce (sometimes significantly) the size of the DTs, the effect is generally less spectacular.
- The DTs corresponding to $\alpha = 0.00005$ can be two to ten times smaller than those corresponding to $\alpha = 1.0$; besides:
- When α decreases from 0.005 to 0.00005, the accuracy of the DTs varies very slowly, and generally insignificantly (cf. the statistical uncertainty of R^b).

FIG. 6. Influence of α on the normalized test set error estimate $N = 500$, $M = 1500$

Figures 5 and 6 illustrate graphically the above observations in the two-, three-, and four-class situations. In each class, the curves represent the mean relative variations of the size, (N_t), and of the misclassification rate (R^b), of 132 DTs (each point of the curves represents the mean value of 12 different DTs built for the corresponding value of α).

The following example provides an insight into the way the value of α operates in the particular case where the attributes contain less information (i.e. the "less separable" case). This corresponds to a LS of 500 states, a biclass pattern, and seven values of α ranging from 1.0 (the nodes are split until they are all leaves) to 0.00005 (extremely "cautious" behavior). Seven DTs were accordingly built, for the contingency #2, where the most important attribute (PG122, the active power generated at the bus #112) was removed from the list of candidate attributes. Table 3 summarizes the results. Columns 2–5 list the number of respectively test nodes, leaves, deadends, total. (Note that since the DTs are binary, they always have as many nodes as the double of test nodes plus one: $N_t = 2N_k + 1$). The following three columns of Table 3 provide the misclassification rate as appraised by respectively the resubstitution, the test sample, and the cross-validation estimate; incidentally, observe the optimistic character of R^b , and to a much lesser extent, of R^{cv} .

One can see that decreasing the value of α in the interval [1.0...0.01] not only drastically reduces the complexity of the DTs but it moreover significantly improves their accuracy. On the other hand, in the interval [0.01...0.00005] the size of the DTs decreases moderately, whereas their accuracy remains unchanged. Figure 7 is an eloquent illustration of the decrease in complexity: for $\alpha = 1$, N_t amounts to 63, whereas for $\alpha = 0.0001$ N_t reduces to 5. Notice that the two DTs have the same structure nearby their respective roots.

The general conclusions are the following:

1. The statistical hypothesis test is able to detect and identify the deadends in a very efficient and

reliable manner, provided that the value of α is lower than 0.001;

2. Using α values below 0.001 provides the twofold benefit of reduced complexity and improved reliability; this effect is even more important in the less separable cases, where the "variance" effect can be very important;

3. The precise value of α , realizing the best compromise between what are called "variance" and "bias" in Section 2, lies somewhere in between 0.001 and 0.00005; the lower the number of classes, the lower the "optimal" value; moreover, the higher the contribution of the variance effect (e.g. the lesser the information contained in the candidate attributes), the lower the optimal value of α ;

4. Anyhow, the bias effect remains very low (it would probably appear markedly for values of α much lower than 0.00005); thus the precise value of α is practically of no concern, as long as it lies in the range [0.001...0.00005];

5. Hence, considering that for a required accuracy, the smaller the trees the better, it is advisable to use $\alpha = 0.0001$ in all cases; sometimes it is even preferable to sacrifice a little accuracy for simplicity of the tree structure.

Remark. Although the above conclusions are drawn in the specific context of transient stability, preliminary investigations indicate that they correspond to the very nature of the statistical hypothesis test, and should remain valid in general. (Wehenkel, 1990).

4.3.3. *On the size of the LS.* One may distinguish the following three questions:

How does the size of the LS influence a DT's complexity and accuracy?

What is the minimal number of learning states required to achieve an acceptable accuracy?

How "stable" is the DT when the LS changes? or stated otherwise, how sensitive is the structure of the DT to the changes of the LS?

Qualitatively, according to the principle of the inductive inference method, one can say that, for a given value of α , increasing the size of the LS will generally (although not necessarily) increase the complexity and the accuracy.

Quantitatively, the answer to the first two questions strongly depends on the particular application of concern and especially on the intricacy of the underlying relationship between the attributes and the classification pattern. The more intricate this relation, the larger the sufficiently accurate DTs and the larger the LS required to build them in a reliable fashion.

As regards stability, we generally found that the nodes near the root of the DT (which

TABLE 3. IMPACT OF α ON COMPLEXITY AND ACCURACY IN THE LESS SEPARABLE CASE OF A BICLASS PATTERN $N = 500$, $M = 1500$, CONTINGENCY #2

α	N_k	N_l	N_{od}	N_t	$R^b(\%)$	$R^s(\%)$	$R^{cv}(\%)$
1.00000	31	32	0	63	0.6	11.9	10.2
0.10000	27	26	2	55	1.0	11.5	7.4
0.01000	3	3	1	7	6.0	8.8	7.4
0.00500	3	3	1	7	6.0	8.8	7.4
0.00050	3	3	1	7	6.0	8.8	7.4
0.00010	2	2	1	5	6.0	8.8	7.2
0.00005	2	2	1	5	6.0	8.8	7.2

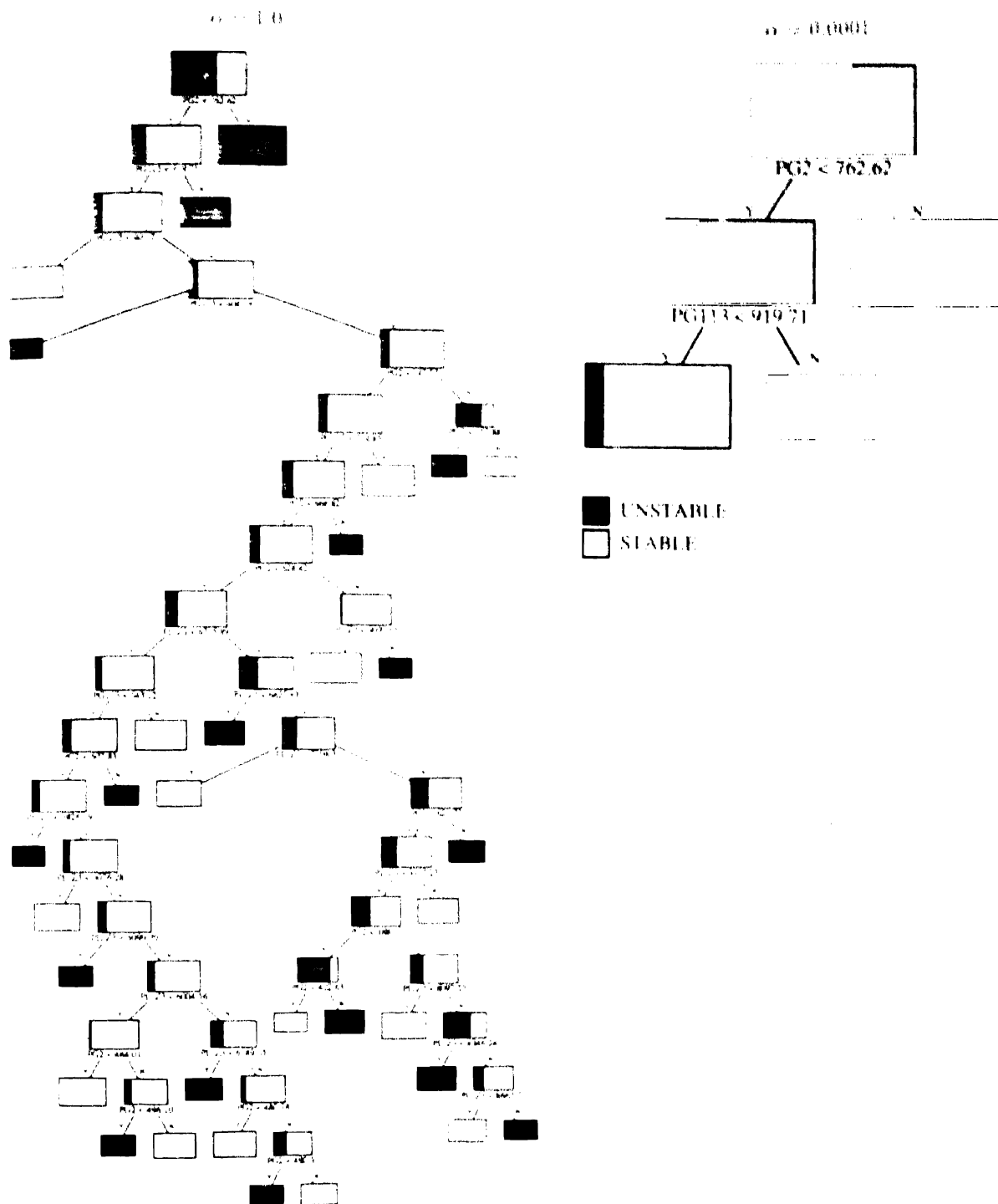


Fig. 7. Typical variation of tree structure with α (biclass DTs for contingency #2, $N = 500$)

express the most relevant relationships) do not change much when N increases, whereas the lower nodes (expressing the details) can change more significantly.

To illustrate and support these considerations, we conducted the following investigations: for the contingency #21, $\alpha = 0.00005$, and in the four-class pattern, six DTs were built with N , the size of the learning set, varying from 100 to 1250 states. Each DT was tested on the basis of the remaining $M = 2000 - N$ test states. The results

are represented graphically in Figs. 8 and 9. One can see that for increasing values of N the error rate decreases from 12.2% to 5.5%, whereas the number of the trees' nodes increases from 9 to 43. At the same time, the number of *retained* attributes is found to increase from 2 to 11. Figure 10 provides two DTs, built for $N = 250$ and 750, and having respectively $N_1 = 9$ and 25, and $A = 2$ and 7. Observe the stability of the DTs.

These and many other similar results indicate

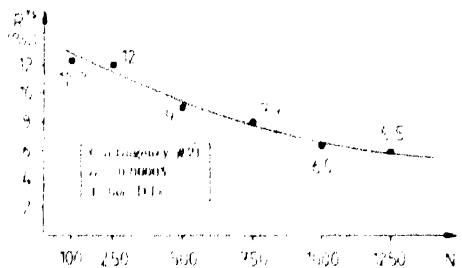


FIG. 8. Influence of the LS size, N , on the test set error estimate.

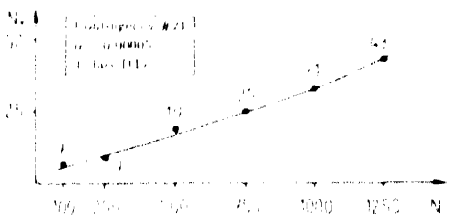


FIG. 9. Influence of the LS size, N , on the number of nodes.

that a size of 500 learning states seems to be a good compromise, allowing the building of sufficiently reliable and moderately complex DTs.

4.3.4. *Candidate attributes.* What would happen if the most relevant attribute (i.e. the one selected for the test at the tree root) were removed from the list of candidate attributes? and what if additional candidate attributes were provided to the procedure?

Removing from the candidate list the most significant attribute, causes a decrease in the accuracy of the DT; however, if “good” alternative attributes remain in the list, this degradation is rather restricted. A ground of comparison has been given by Tables 2 and 3. In Table 2 a DT composed of 7 nodes was obtained

for the two-class case corresponding to contingency #2, when a sufficiently complete list of candidate attributes is used (including in particular the most relevant attribute for this case, namely PG_{112} , the active power generated at bus #112). Table 3 indicates the effect of removing PG_{112} from the list of candidate attributes: the tree corresponding to $\alpha = 0.0001$ reduces to 5 nodes and its probability of misclassification R^b increases from 2.27% to 8.80%. This is further illustrated in Fig. 11, where the two biclass trees are presented: obviously, removing the best attribute, has caused a degradation of the tree’s accuracy which, nevertheless, remains quite good (8.80% vs 2.27%).

Conversely, providing additional relevant attributes will generally improve the DT accuracy; however, if the most significant ones are already included in the list, only the lower parts of the DT will be affected, and the increase in accuracy will be almost negligible.

As a conclusion, if the important attributes are not known *a priori*, it is advisable to use an as large as possible list of candidate attributes. The first constructed DT identifies itself the relevant attributes, which can be used for subsequent tree constructions, possibly in addition to new ones.

4.3.5. *Misclassification errors as related to the number of classes.* In Section 4.3.1 we observed that the larger the number of a tree classes, and the larger its complexity and misclassification error rate. This is not surprising though: for a given LS, the more detailed the information one wants to extract, the larger its inaccuracy; as a counterpart, one may reasonably expect this inaccuracy to be less harmful.

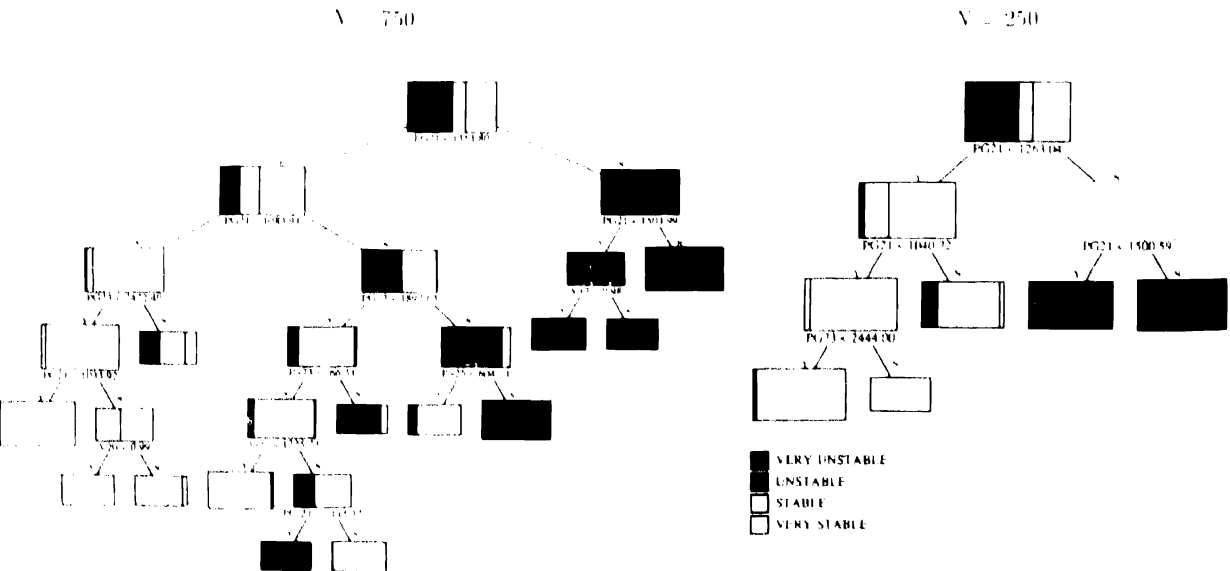


FIG. 10. Influence of the LS size, N , on the tree structure (contingency #21, $\alpha = 0.00005$).

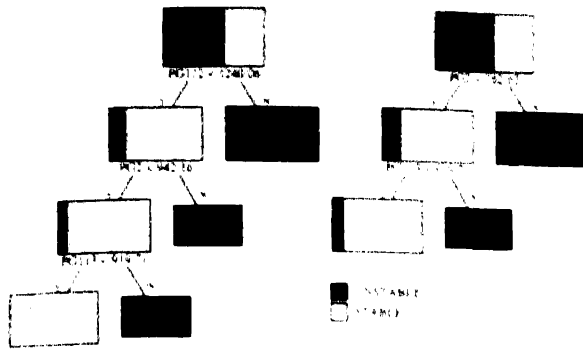


FIG. 11 Effect of removing the best attribute PG_{12} (contingency #2, $\alpha = 0.00005$, $N = 500$)

All our investigations corroborate this reasoning. A typical case is considered in Tables 4 and 5, which collect data for respectively a two- and a three-class tree, built for the contingency #21, $\alpha = 0.00005$, $N = 500$ and $M = 1500$. The true classification of these 1500 test states is reported in the columns of the tables, where the labels VU, U, S, VS stand for respectively very unstable, unstable, stable, very stable. Their classification as provided by the (biclass and four-class) trees, is reported in the rows of the tables; an additional row provides the misclassification rates. These latter are overall quite

TABLE 4 TYPICAL MISCLASSIFICATION ERRORS OF A TWO-CLASS TREE ($\alpha = 0.00005$, $N = 500$, $M = 1500$)

		True classification		
		U	S	All
Tree classification	U	723	25	748
	S	56	696	752
	All	779	721	1500
Errors (%)		3.73	1.67	5.4

TABLE 5 TYPICAL MISCLASSIFICATION ERROR DISTRIBUTION OF A FOUR-CLASS TREE ($\alpha = 0.00005$, $N = 500$, $M = 1500$)

		True classification				
		VU	U	S	VS	All
Tree classification	VU	461	29	0	0	490
	U	8	217	8	0	233
	S	7	56	189	8	260
	VS	0	1	20	496	517
All		476	303	217	504	1500
Errors (%)		1.9	5.73	1.82	0.53	9.13

reasonable, as is also suggested by the strongly diagonal dominant character of the "kernel" of the tables.

Observe also that the total error of the biclass tree is lower than of the four-class tree; but the latter error is less misleading than the former; for example, declaring unstable a state which is actually stable is less misleading in the four-class tree, because of the finer definition of the four classes. Stated otherwise, in the four-class tree, a large majority of errors appear among neighboring classes (e.g. see Fig. 12): the error distribution concentrates mainly around the true class and the number of "outliers" almost (if not totally) reduces to zero. Observe also that a finer exploration and identification of the misclassified states indicates that in the biclass tree these latter mainly concentrate in the vicinity of the stable-unstable borderline as well; hence their misclassification is not totally misleading, after all.

4.3.6 On the accuracy of the R^b and R^c estimates. To quantify the considerations of Section 2.8, we have compared the accuracy of the R^b and R^c estimates with respect to R^a considered as the benchmark, because of its high reliability.

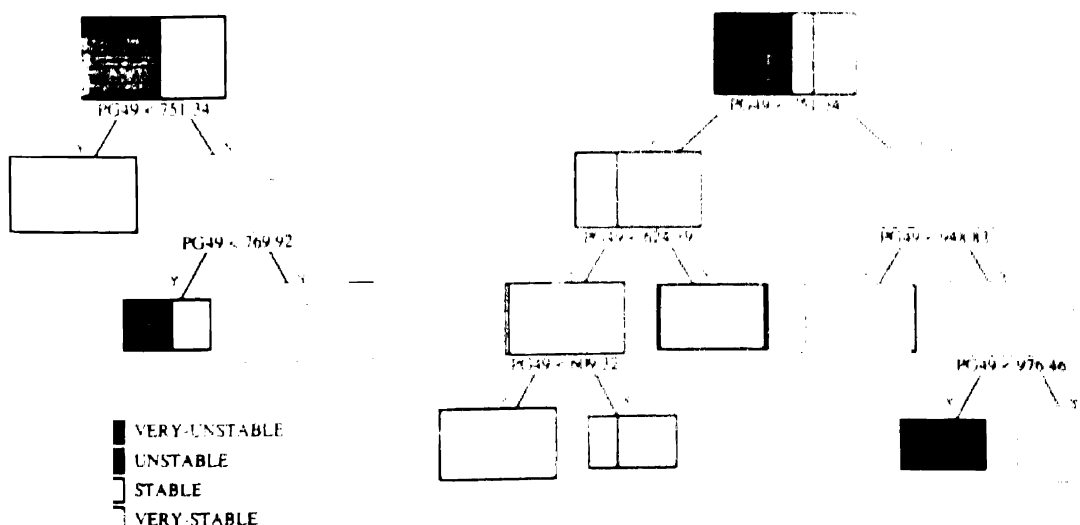


FIG. 12. Typical bi- and four-class trees (contingency #49, $\alpha = 0.00005$, $N = 500$)

- With respect to R^b , we observe the following:
- In all cases, it is strongly optimistically biased;
 - The bias varies from almost 100% underestimation for the high values of α to approximatively 40% underestimation for $\alpha = 0.00005$;
 - For a fixed value of α , the bias decreases slowly when the number N of learning states increases;
 - The bias could become acceptable only for values of α much lower than 0.00005;
 - Thus, for "good" values of α resulting from our earlier discussion, this measure is of no practical value for the estimation and comparison of accuracies.

Regarding the accuracy of R^{cv} , we observe the following:

- For a given DT, its precise value is almost independent of the value of the parameter V , as long as this value is larger than 10.
- R^{cv} sometimes overestimates, sometimes underestimates the DTs actual accuracy, depending on the particular DT of concern;
- Defining the error of R^{cv} by $E^{cv} \triangleq R^{cv} - R^b$ we found that the mean value of E^{cv} over 14 different DTs and for different values of V , is -0.17% , whereas the standard deviation of E^{cv} is 2.15% (for the purpose of comparison we mention that the variance of R^b , is about 0.7% for $M = 1500$, and about 2% for $M = 150$);
- This corroborates our earlier statement that the variance of R^{cv} is rather high;
- As a conclusion, this estimate is particularly interesting when the total number of observations $(N + M)$ is too small in practice to be split into an LS and a sufficiently large $(M > 500)$ test set.

4.4. A global set of simulations

Another set of simulations concerning a less refined but more global investigation than that of the previous Section 4.3 has been carried out using the same data base of 2000 OPs: 31 contingencies of the 3 ϕ SC type, successively applied at each generator bus of the 31-machine system which have been pre-analysed via the extended equal area criterion. Table 6 collects information for the two-, three-, and four-class patterns (N_i and A stand for the total number of nodes and of selected attributes). Its last line summarises the mean characteristics of the DTs in terms of the number of stability classes.

Note that their complexity and accuracy depend almost linearly on the number of classes. Observe also that in terms of accuracy, this global assessment is certainly pessimistic, unfavorable to the method. This is due to the fact

TABLE 6. GLOBAL ASSESSMENT 3 ϕ SC APPLIED AT EACH GENERATOR BUS $\alpha = 0.0001$, $N = 500$, $M = 1500$

Gen. Bus #	Two classes			Three classes			Four classes		
	N_i	$R^b(\%)$	A	N_i	$R^b(\%)$	A	N_i	$R^b(\%)$	A
2	7	2.67	3	19	5.40	5	25	9.53	6
11	7	3.73	2	17	6.67	6	21	11.40	4
19	5	4.33	2	13	11.73	2	27	11.87	8
21	5	4.00	1	17	5.60	5	19	8.13	5
22	5	3.93	1	17	7.13	5	21	10.47	3
29	3	2.47	1	9	4.33	2	15	7.87	4
30	5	3.93	2	17	8.07	6	23	9.60	6
39	3	2.07	1	7	4.27	1	9	5.60	2
41	7	3.87	2	19	13.20	5	25	12.67	6
44	9	2.73	2	13	6.20	5	15	9.20	3
48	15	4.67	4	23	8.27	7	27	15.87	6
49	3	1.93	1	7	5.07	1	13	9.47	2
50	15	3.73	5	23	9.13	7	29	11.80	9
59	7	5.13	3	17	7.33	5	23	12.33	5
65	5	2.27	1	15	3.40	3	11	6.73	2
66	3	1.67	1	9	3.40	3	17	6.33	4
70	7	3.00	2	11	7.93	3	19	7.93	5
71	7	3.27	2	15	8.40	5	21	12.67	6
72	5	2.20	2	9	5.33	3	11	9.20	3
73	3	2.73	1	11	5.13	3	19	7.87	4
74	17	4.20	6	23	8.87	6	31	17.07	9
75	5	3.13	1	11	3.87	2	19	10.27	6
79	3	2.00	1	9	4.80	2	15	7.40	4
84	9	4.33	3	13	8.20	3	25	10.40	7
95	5	2.00	2	15	6.13	4	21	8.67	4
112	9	1.93	3	15	5.00	3	15	5.40	3
113	5	5.00	2	15	8.67	5	29	11.40	8
123	11	5.60	4	9	8.13	3	31	9.73	9
124	7	2.53	2	15	3.80	4	15	9.47	3
129	7	4.00	2	19	9.47	5	17	13.53	6
132	7	1.73	2	7	5.80	1	13	8.47	3

Mean values over 31 contingencies

6.8 3.25 2.2 14.2 6.72 3.9 20 9.95 5.0

that the candidate attributes used to construct the trees have been chosen so as to cover the whole power system; no particular care has been taken to consider a more refined list of attributes around the contingencies locations. (Remember that for the 31 contingencies, only 83 attributes are used, often not close enough to some of them.) Another, although less important source of inaccuracy is the fact that the CCTs used for classifying the OPs of both the LS and the TS have been computed via the EEAC and not the numerical integration method; this, unavoidably introduces a small bias in most cases. More detailed information may be found in Wehenkel (1990).

5. DISCUSSION

The questions posed and the answers given in the preceding sections are reorganized so as to draw general conclusions. Some pertain specifically to the DTTS method, some others apply to the DT methodology in general.

The statistical test used to stop splitting the nodes of a grown tree appears to work very

satisfactorily. Its threshold α provides an effective tool for controlling the complexity and the accuracy of the resulting tree. It is particularly interesting that simplicity and accuracy of a DT are not contradictory objectives, at least in a rather large range of the α values. Indeed, in this range, decreasing α does not significantly affect the accuracy, whereas it contributes to drastically decrease the number of nodes of the tree. Incidentally, this explains *a posteriori* the good performances obtained in our earlier investigations where α was fixed rather arbitrarily to 0.01. Overall, the simplicity and accuracy of the trees provided by the method are quite remarkable. This seems to be a general feature of the inductive inference method we developed to build DTs.

The particular DTTS approach proves also to be very effective in many respects, and the underlying conjectures fully justified. For example, among the large number of attributes proposed to the method, only a few are retained as the relevant parameters driving the transient stability phenomena. Moreover, increasing the number of classification patterns of a tree provides additional relevant attributes: this allows to get a more refined insight into the mechanism of the phenomena, and to offer additional means to control transient stability. The interplay between biclass trees—with extremely simple structures and reduced number of relevant attributes, and multiclass trees—with more complex (yet tractable) structures and larger (yet restricted) number of attributes, is another attractive feature of the method, which thus shows to be very flexible and stable. This stability of a tree with respect to the attributes is a very interesting aspect, indeed: it amounts to systematically using the same, more relevant attributes nearby its root, whatever the number of its classes and its complexity.

The tradeoff between bi- and multiclass DTs appears thus to be a great asset of the DTTS method, not a drawback. Various solutions, supplementing different, complementary information extracted from a LS, may thus be exploited for various purposes, even at the price of a somewhat lesser accuracy in the multiclass case.

Investigations relating to the construction of an "adequate" LS have pointed out another interesting aspect, namely that reasonable sizes of LSs are sufficient to build reliable DTs. On the other hand, the states composing the LSs were chosen on a statistical basis, the purpose sought here being the objective assessment of the DTTS method. Note that in a real world context, this choice should take into account requirements imposed by the power system of

concern, specified in collaboration with the engineers and operators in charge of the system. Reconsidering the "right size" of a LS, one may wonder to which extent this should depend on the size of the power system. To answer this question, one probably should specify whether the purpose is to build trees for contingencies spread throughout the whole system, or whether one seeks to explore some particular contingencies. In the former case, a LS covering the whole power system with sufficiently detailed information, would indeed be needed, its size would therefore increase with that of the power system. In the latter case, the LS should essentially contain detailed information only for the regions of concern for these contingencies.

One could object that the above conclusions rely on the particular, very severe type of contingencies considered so far, and also on particularly simplified system modelling. However, our objective was the validation of the method as such, we believe that this objective has been encountered.

Certainly, to be interesting a method has to be computationally tractable. The considerations of Section 3.1.4 indicate that the main, and in fact sole, really burdensome task is the construction of the data base. But this has to be done only once, then occasionally updated. The construction of the DTs, although also off-line, is a less demanding task; and the better the knowledge of the power system, the faster the construction of its trees. Once constructed, the storage and use of the DTs are extremely inexpensive, thousands of DTs could be simultaneously stored into main memory, as for the mean classification time, it was assessed to be about 0.6 ms, i.e. almost negligible.

The way of using the DTs was not considered in this paper, only classes of possible uses were enumerated, and prospects for their real-time applications to transient stability analysis, sensitivity analysis and preventive control were suggested. Nevertheless, other interesting applications of the DTTS approach may be foreseen as well, in the context of training simulators, and of planning studies.

6. CONCLUSION

Two main objectives have been pursued in this paper. First, to get in-depth knowledge of the inductive inference method designed in our previous studies, and more specifically of its stop splitting criterion. Second, to scrutinize the basic features of the decision tree transient stability (DTTS) method, i.e. of the inductive inference method as applied to transient stability of power systems.

To encounter these objectives, a large-scale investigation was conducted, using a realistic power system. The obtained results are quite interesting. As regards the inductive inference method, it was proven to be very efficient, indeed, appropriate to yield simple and accurate DTs; and although it would be hazardous to compare methods used in different application domains, one nevertheless may say that it appears to be among the very effective methods reported in the technical literature.

Concerning its application to transient stability, the devised DTTS method has exhibited very attractive features, with manifold potential. For example, it was found to be capable of treating the three aspects of transient stability assessment, viz. analysis, sensitivity, and control. It could therefore be exploited in planning studies. At least as interesting are its real-time aspects, and its capabilities in on-line transient stability assessment and preventive control.

Admittedly, many other questions still remain unexplored. This study was the first, indispensable step towards real world application of the DTTS method.

REFERENCES

- Bergen, A. R. (1986). *Power System Analysis*. Prentice Hall, Englewood Cliffs, NJ.
- Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984). *Classification and Regression Trees*. Belmont, Wadsworth.
- EPRI EL-4958 Project 2496-1 (1987). Dynamic Security Assessment for Power Systems: Research Plan. Final Report.
- Friedman, J. H. (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Trans. Computers*, **C-26**, 404-408.
- Kononenko, L., I. Bratko and I. Roskar (1984). Experiments in automatic learning of medical diagnosis rules. Technical Report. Jozef Stefan Institute, Ljubljana, Yugoslavia.
- Kvålseth, T. O. (1987). Entropy and correlation: some comments. *IEEE Trans. Syst., Man Cybern.*, **SMC-17**, 517-519.
- Lee, S. T. Y. (1972). *Transient stability equivalents for power system planning*. Ph.D. Thesis, Massachusetts Institute of Technology, MA.
- Quinlan, J. R. (1984). Learning efficient classification procedures and their application to chess endgames. In R. S. Michalski, J. G. Carbonell and T. M. Mitchell. *Machine Learning: An Artificial Intelligence Approach*, pp. 463-482. Springer, Berlin.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, **1**, 81-106.
- Ribbens-Pavella, M., and F. J. Evans (1985). Direct methods for studying dynamics of large scale electric power systems—A survey. *Automatica*, **21**, 1, 1-21.
- Toussaint, G. T. (1974). Bibliography on estimation of misclassification. *IEEE Trans. Inf. Theory*, **IT-20**, 472-479.
- Wehenkel, L., Th. Van Cutsem and M. Ribbens-Pavella (1986). Artificial intelligence applied to on-line transient stability assessment of electric power systems. *Proc. 25th IEEE Conf. Decision and Control*, pp. 649-650. Athens, Greece.
- Wehenkel, L., Th. Van Cutsem and M. Ribbens-Pavella (1987). Artificial intelligence applied to on-line transient stability assessment of electric power systems. *Proc. 10th IFAC World Congress*, pp. 308-313, Munich, F.R.G.
- Wehenkel, L., Th. Van Cutsem and M. Ribbens-Pavella (1988). Decision trees applied to on-line transient stability assessment of power systems. *Proc. IEEE Int. Symp. on Circuits and Systems*, Vol. 2, pp. 1887-1890, Helsinki, Finland.
- Wehenkel, L. (1988). Artificial intelligence methods for on-line transient stability assessment of electric power systems. *Proc. Symp. on Expert Systems Application to Power Systems*, pp. 5.1-5.8, Stockholm-Helsinki.
- Wehenkel, L., Th. Van Cutsem and M. Ribbens-Pavella (1989a). Inductive inference applied to on-line transient stability assessment of electric power systems. *Automatica*, **25**, 445-451.
- Wehenkel, L., Th. Van Cutsem and M. Ribbens-Pavella (1989b). An artificial intelligence framework for on-line transient stability assessment of power systems. *IEEE Transactions Power Systems*, **PWRS-4**, 789-800.
- Wehenkel, L. (1990). *Une Approche de l'Intelligence Artificielle Appliquée à l'Évaluation de la Stabilité Transitoire des Réseaux Électriques*. Ph.D. Thesis (in French), University of Liège, Belgium.
- Xue, Y., Th. Van Cutsem and M. Ribbens-Pavella (1988). A simple direct method for fast transient stability assessment of large power systems. *IEEE Trans. Power Systems*, **PWRS-3**, 400-421.

Brief Paper

One-step Optimal Saturation Correction*

N. L. SEGALL,† J. F. MACGREGOR,‡§ and J. D. WRIGHT,‡

Key Words—Control system design, discrete time systems, optimal control, saturation, linear optimal regulator.

Abstract—A one-step optimal method for compensating for any form of input saturation in discrete linear controllers is developed. The algorithm is straightforward to implement as an analytical correction operating on the difference between the past calculated control inputs and the inputs actually implemented. A further correction for multivariable controllers is necessary to simultaneously adjust the remaining control inputs if some inputs saturate. The algorithm is illustrated by simulation on several SISO and MIMO examples.

1. Introduction

ONE OF THE most common problems encountered in control of process systems is that the control actions calculated by a controller cannot be implemented in full, that is, the control input saturates. Linear controllers only meet their design specifications when the system is in its linearized operating region; if a calculated control input to the system violates a saturation limit then the subsequent control will not in general be satisfactory.

Goodwin (1972) presented a one-step optimal saturation compensation algorithm for single-input single-output minimum variance controllers. Benzanson (1984) extended this to more general SISO one-step optimal LOG controllers. A different method of implementing one-step optimal controllers for SISO systems was presented by Clarke (1981). Makila (1982) attempted to implement the result of Goodwin (1972) and to extend it to a multivariable one-step controller. Toivonen (1983a, b, c) presented an approximation to the multistep LOG problem with saturation of the form $A^{-1}U_1 \leq B$, and Parrish and Brosilow (1985) presented some empirical rules for saturation compensation for what they refer to as inferential controllers.

This paper presents a new one-step optimal correction for any form of input saturation in multivariable controllers. A further simultaneous correction is developed which causes the remaining control inputs to compensate for saturation in other inputs.

2. The algorithm

This section presents the derivation of a controller which minimizes the one-step objective function

$$J_1 = E\{Y_{t+k}^T Q Y_{t+k} + U_t^T \nabla^d R \nabla^d U_t + Y_t^T V_t + U_t^T U_t\} \quad (1)$$

subject to any set of time varying constraints on the control vector U_t . Q and R are $n \times n$ weighting matrices and Q is positive definite. The control algorithm can easily be generalized to include more general polynomial matrices operating on Y_{t+k} and U_t in the objective function (1).

This minimization is subject to the condition that past inputs (U_t) may, for any reason whatsoever, have not been implemented as called for. No assumption is made in this derivation about the form of the saturation limits on the past U_t . The derivation follows the single variable derivations of Clarke and Hastings-James (1971) and Clarke and Gawthrop (1975), and the multivariable derivation of Koivo (1980) for the case of no hard constraints.

The system is assumed to be a linear multivariable system which is modelled in the ARMAX form,

$$A(q^{-1})Y_t = B(q^{-1})U_t + C(q^{-1})e_t \quad (2)$$

where Y_t is an n -dimensional vector of process output deviations from the setpoint vector Y_{sp} , or from steady state, U_t is an n -dimensional vector of process input deviations from steady state, k is the number of whole periods of process delay, and e_t is an n -dimensional vector of zero mean white noise processes. A , B and C are $n \times n$ polynomial matrices in the backward shift operator q^{-1} . B_0 is assumed nonsingular and $A_0 = C_0 = I$. The roots of the determinants of the $A(q^{-1})$ and $C(q^{-1})$ polynomial matrices all have their roots on or inside the unit circle in q . The $A(q^{-1})$ and $B(q^{-1})$ matrices may have a common scalar factor ∇^d where ∇ is the backwards difference operator $1 - q^{-1}$. This allows for nonstationarity of the disturbance, and d (usually zero or one) is the degree of nonstationarity. MacGregor *et al.* (1984) show the duality between ARMAX models for stochastic and deterministic disturbances and for setpoint changes, so the system model may be selected for regulation or for servo control.

Substituting the model (2) into the objective function (1) gives

$$\begin{aligned} \text{Min } E\{[A^{-1}(q^{-1})B(q^{-1})U_t + A^{-1}(q^{-1})C(q^{-1})e_{t+k}]^T Q \\ \times [A^{-1}(q^{-1})B(q^{-1})U_t + A^{-1}(q^{-1})C(q^{-1})e_{t+k}] \\ + U_t^T \nabla^d R \nabla^d U_t\} \end{aligned} \quad (3)$$

The effect of the disturbance $A^{-1}(q^{-1})C(q^{-1})e_{t+k}$ in equation (3) can be separated into forecastable and unforecastable parts by solving the linear Diophantine equation

$$C(q^{-1}) = A(q^{-1})F(q^{-1}) + q^{-k}G(q^{-1}) \quad (4)$$

for the least degree solution with respect to $F(q^{-1})$. The degree of $F(q^{-1})$ will be k and $F_0 = I$. Substituting for

* Received 6 November 1986; revised 30 October 1987; revised 23 May 1989; received in final form 4 May 1990. The original version of this paper was presented at the 10th IFAC World Congress which was held in Munich, F.R.G. during July, 1987. The published Proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak.

† Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L7.

‡ Xerox Research Centre of Canada, 2660 Speakman Drive, Mississauga, Ontario, Canada L5K 2L1.

§ Author to whom all correspondence should be addressed.

$\mathbf{A}^{-1}(q^{-1})\mathbf{C}(q^{-1})$ in (3) using (4), and taking conditional expectations gives

$$\begin{aligned} \min_{U_t} \{ & [\mathbf{A}^{-1}(q^{-1})\mathbf{B}(q^{-1})\mathbf{U}_t + \mathbf{A}^{-1}(q^{-1})\mathbf{G}(q^{-1})\mathbf{e}_t]^T \mathbf{Q} \\ & \times [\mathbf{A}^{-1}(q^{-1})\mathbf{B}(q^{-1})\mathbf{U}_t + \mathbf{A}^{-1}(q^{-1})\mathbf{G}(q^{-1})\mathbf{e}_t] \\ & + \mathbf{U}_t^T \nabla^d \mathbf{R} \nabla^d \mathbf{U}_t \}. \end{aligned} \quad (5)$$

To minimize the quadratic cost function (5) with respect to only the current control action \mathbf{U}_t , expression (5) is differentiated with respect to \mathbf{U}_t and the result is set to zero giving

$$\begin{aligned} \mathbf{J}_{t,u} = & 2\mathbf{B}_0^T \mathbf{Q} [\mathbf{A}^{-1}(q^{-1})\mathbf{B}(q^{-1})\mathbf{U}_t \\ & + \mathbf{A}^{-1}(q^{-1})\mathbf{G}(q^{-1})\mathbf{e}_t] + 2\mathbf{R} \nabla^d \mathbf{U}_t, \end{aligned} \quad (6)$$

where $\mathbf{J}_{t,u}$ is the derivative of the objective function with respect to \mathbf{U}_t . Eliminating \mathbf{e}_t using (2) and simplifying using (4) gives

$$\begin{aligned} \mathbf{J}_{t,u} = & 2\mathbf{B}_0^T \mathbf{Q} \mathbf{A}^{-1}(q^{-1})\mathbf{G}(q^{-1})^T \mathbf{C}^{-1}(q^{-1})\mathbf{A}(q^{-1})\mathbf{Y}_t \\ & + 2\mathbf{B}_0^T \mathbf{Q} \mathbf{F}(q^{-1})\mathbf{C}^{-1}(q^{-1})\mathbf{B}(q^{-1})\mathbf{U}_t + 2\mathbf{R} \nabla^d \mathbf{U}_t. \end{aligned} \quad (7)$$

Equation (7) is now simplified using the relationship (Åström, 1978)

$$\begin{aligned} \mathbf{C}(q^{-1})\mathbf{F}^{-1}(q^{-1})\mathbf{A}^{-1}(q^{-1})\mathbf{G}(q^{-1}) \\ = \mathbf{G}(q^{-1})\mathbf{F}^{-1}(q^{-1})\mathbf{A}^{-1}(q^{-1})\mathbf{C}(q^{-1}) \end{aligned} \quad (8)$$

giving

$$\begin{aligned} \mathbf{C}(q^{-1})\mathbf{F}^{-1}(q^{-1})\mathbf{Q}^{-1}\mathbf{B}_0^{-1}\mathbf{J}_{t,u} = & 2\mathbf{G}(q^{-1})\mathbf{F}^{-1}(q^{-1})\mathbf{Y}_t \\ & + 2\mathbf{B}(q^{-1})\mathbf{U}_t + 2\mathbf{C}(q^{-1})\mathbf{F}^{-1}(q^{-1})\mathbf{Q}^{-1}\mathbf{B}_0^{-1}\mathbf{R} \nabla^d \mathbf{U}_t. \end{aligned} \quad (9)$$

The current control action is one-step optimal if it is chosen so that $\mathbf{J}_{t,u}$ is zero. Define \mathbf{U}_t^* as the control action which would set $\mathbf{J}_{t,u}$ to zero in the absence of any saturation limits, and \mathbf{U}_t as the control action that is actually implemented; that is,

$$\mathbf{U}_t = \text{sat} \{ \mathbf{U}_t^* \} = \begin{cases} \mathbf{U}_t^* & \text{if no limit is encountered} \\ \mathbf{U}_{\text{limit}} & \text{if a limit is encountered} \end{cases} \quad (10)$$

If the calculated control action cannot be implemented due to saturation in any form then $\mathbf{J}_{t,u}$ will not be zero. In this case it can be seen from (9) that $\mathbf{J}_{t,u}$ will be given by

$$\mathbf{J}_{t,u} = 2[\mathbf{B}_0^T \mathbf{Q} \mathbf{B}_0 + \mathbf{R}] \mathbf{U}_t^* \quad (11)$$

where \mathbf{U}_t^* is the unimplemented portion of the control action called for at time t , that is

$$\mathbf{U}_t^* = \mathbf{U}_t^* - \mathbf{U}_t \quad (12)$$

So, in solving for the control actions which set $\mathbf{J}_{t,u}$ to zero, the past values of $\mathbf{J}_{t,u}$ may not be zero and are defined by equation (11)

To produce a controller which can be implemented express $\mathbf{F}^{-1}(q^{-1})$ as the adjoint $\mathbf{F}^*(q^{-1})$ divided by the determinant $|\mathbf{F}(q^{-1})|$. The determinant is a scalar polynomial and thus it can multiply equation (9). Setting $\mathbf{J}_{t,u} = 0$ and substituting for past values of $\mathbf{J}_{t,u}$ using equation (11) gives

$$\begin{aligned} \mathbf{G}(q^{-1})\mathbf{F}^*(q^{-1})\mathbf{Y}_t + |\mathbf{F}(q^{-1})|\mathbf{B}(q^{-1})\mathbf{U}_t \\ + \mathbf{C}(q^{-1})\mathbf{F}^*(q^{-1})\mathbf{Q}^{-1}\mathbf{B}_0^{-1}\mathbf{R} \nabla^d \mathbf{U}_t \\ + [|\mathbf{I} - \mathbf{C}(q^{-1})\mathbf{F}^*(q^{-1})|]\mathbf{B}_0 + \mathbf{Q}^{-1}\mathbf{B}_0^{-1}\mathbf{R} \mathbf{U}_t^* = 0 \end{aligned} \quad (13)$$

Equation (13) can now be used to calculate the control action \mathbf{U}_t^* which would minimize (1) if it could be implemented, that is, if no saturation limits are violated (it calculates the globally optimal input which may or may not be feasible). The actually implemented control action can be calculated if the nature of the saturation limits are known, or may be obtained by direct feedback from the actuator if this is available.

If the system being controlled is a single-input single-output system then the optimal control input subject to the saturation limits is either that calculated from equation (13) or the limit if a saturation limit is violated. For SISO system $\mathbf{F}^*(q^{-1}) = 1$ and $|\mathbf{F}(q^{-1})| = \mathbf{F}(q^{-1})$. If the system being controlled is a multi-input multi-output system then a

further correction is required to find the current one-step optimal input subject to the saturation limits. A simultaneous correction for the case where the saturation limits on the current input are of the time varying form $\mathbf{U}_{\text{min},t} \leq \mathbf{U}_t \leq \mathbf{U}_{\text{max},t}$ is derived in Appendix A. The procedure is iterative and converges in at most n steps.

If there is no saturation, then \mathbf{U}_t^* will always be zero and equation (13) defines the same one-step optimal controller as Koivo (1980) and Mäkilä (1982) which is a generalization of the single-input single-output controller of Clarke and Hastings-James (1971) and Clarke and Gawthrop (1975). Equation (7) is a multivariable generalization of the one-step controller with input limits of Goodwin (1972) and Bezanson (1984). It involves rational polynomial matrices in q^{-1} which can be implemented either by expanding as infinite polynomials or by introducing auxiliary states. On the other hand equation (13) presented here allows the controller to be implemented easily using a finite number of past values of \mathbf{Y}_t , \mathbf{U}_t and \mathbf{U}_t^* . Clarke (1981) presents a univariate saturation correction which involves three equations: a prediction, a control calculation and a revised prediction. Mäkilä's (1982) implementation of the saturation correction of Goodwin (1972) failed to rearrange the rational polynomial matrices correctly as shown in this paper. His algorithm is lacking the correction term involving past \mathbf{U}_t 's appearing in equation (13). Toivonen (1983a, b, c) presents a solution to the infinite step optimal control problem with saturation constraints which is based on expectations over all future time. This off-line design does not account for the actual history of saturations which has occurred, nor does it allow for time varying constraints. It essentially results in a reduction of the controller gain so that the saturation limits will not be violated.

There is no guarantee that a one-step optimal controller will be stable, or that the closed loop system using a one-step optimal controller will be stable. For this reason it is necessary to check the poles of the controller and of the closed loop transfer function to ensure stability. The constraint parameter \mathbf{R} can be adjusted to obtain stability. Furthermore, if the control actions are strongly oscillating, velocity limits on the control action can lead to instability. The best policy is to ensure both stability and acceptability of the control actions by an appropriate choice of \mathbf{R} at the design stage before applying the one-step optimal controller.

The controller (13) is shown in Segal *et al.* (1986) to also provide one-step optimal correction for past input saturation for Internal Model Controllers (Garcia and Morari, 1982), Dahlin controllers (Dahlin, 1968), and pole placement controllers.

The controller with saturation correction derived here provides a closed-form solution to the optimization of the one-step objective function. This controller is optimal in a one-step sense: it calculates the input which minimizes the objective function (1), taking into account that the past calculated control actions may not have been applied in full. It does not account for whether any future inputs are likely to violate a limit. However, Wong *et al.* (1987), showed by mathematical programming that optimization with this objective function performed better than any of the anti-reset windup methods which they considered.

3. Single-input-single-output simulation results

This section presents a simulated example of controller saturation to demonstrate the proposed algorithm. The simulation shows the servo response to a step setpoint change.

The system considered is an over-damped continuous second-order process which is made up of two cascaded first-order processes plus dead-time. The gain of the process is 1, the dead-time is 1 minute and the two time constants are 1.44 minutes and 1.09 minutes. With a control interval of 1 minute and a zero-order hold between intervals the ARMAX model with step disturbances or step changes in setpoint is

$$\begin{aligned} (1.0 - 0.9q^{-1} + 0.2q^{-2}) \nabla Y_t = & (0.189 + 0.111q^{-1}) \nabla U_t + \\ & + (1.0 - 0.9q^{-1} + 0.2q^{-2}) e_t. \end{aligned} \quad (14)$$

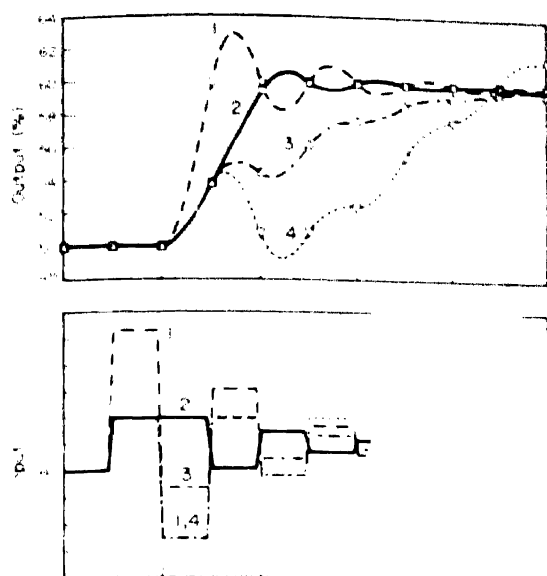


Fig. 1 Dead-beat or minimum variance control for second-order system. Step change in setpoint from 50% to 60% at time 1 minute. 1. No saturation limit. 2, 3, 4. Saturation limit on input at 100%. 2. Saturation correction algorithm used. 3. Reset windup protection only, uses past actually implemented U_i . 4. No saturation correction, uses past calculated U_i .

The controller equations is

$$\begin{aligned} & (0.189 + 0.3q^{-1} + 0.111q^{-2})\nabla U_i \\ & = (1.0 - 0.9q^{-1} + 0.2q^{-2})Y_i \\ & + (0.189)(-0.9q^{-1} + 0.2q^{-2})U_i \quad (15) \end{aligned}$$

The response to a step change at time 1 minute from 50% to 60% in the setpoint is shown in Fig. 1. The initial steady state input is 80%. The figure compares the continuous process response when there is no saturation limit on the input (labelled 1) to the response when there is a saturation limit at 100% and the algorithm presented here is used (labelled 2). In addition, for comparison, the response of the velocity form implementation of the controller which simply prevents reset windup by using the actual past values of the U_i in the calculations is shown (labelled 3), as is the implementation of the controller where saturation is ignored and past values are used as calculated (labelled 4).

4. Multi-input-multi-output simulation results

This section presents a simulated example of controller saturation in a multivariable system to demonstrate both the correction for past saturation and the simultaneous saturation correction. The simulation shows the servo response to a step setpoint change.

The system considered is an experimental two-tank gravity feed process with heaters in each tank and a constant recycle flow from the bottom tank to the top tank. The state space model for the system is (Hugo, 1986):

$$\begin{aligned} \frac{d}{dt} T_T(d) &= -0.0121 T_T(d) + 0.0527 T_B(d) + U_T(d) \\ \frac{d}{dt} T_B(d) &= 0.0208 T_T(d) - 0.0208 T_B(d) + U_B(d) \end{aligned} \quad (16)$$

where T represents the temperature in degrees Celsius and U refers to the heat input in volts at the computer output (range 0–10 V) and the time t is in seconds. The subscripts T and B refer to the top and bottom tanks, respectively. The steady state values are $T_T = 30^\circ\text{C}$, $T_B = 31^\circ\text{C}$, $U_T = 3.0\text{ V}$

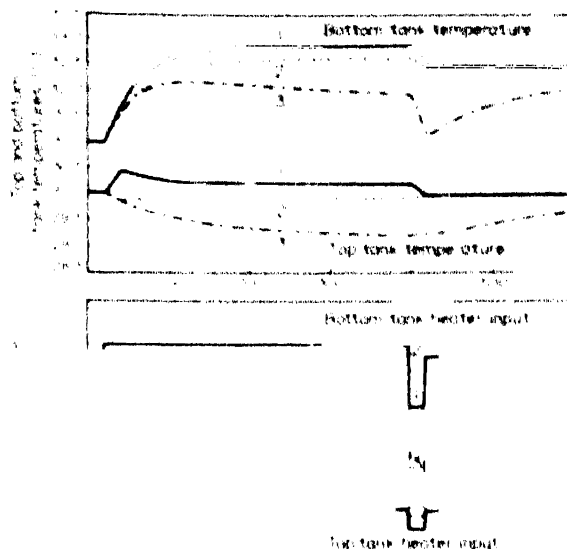


Fig. 2 Multivariable dead-beat or minimum variance control. Step change in bottom tank temperature set point from 31°C to 34°C at time 20 seconds. Step change in bottom tank temperature set point from 34°C to 32.5°C at time 400 seconds. Saturation limits on both inputs at 0 V and 10 V. 1. Saturation correction algorithm with simultaneous correction. 2. Saturation correction without simultaneous correction. 3. Reset windup protection only, uses past actually implemented U_i .

and $U_B = 5.5\text{ V}$. The discrete controller uses a 20 second control interval.

Figure 2 shows the response of the system to a prolonged period of saturation. The setpoint temperature in the bottom tank is changed to the unattainable value of 34°C at time 20 seconds. At time 400 seconds, the setpoint is changed to the attainable value of 32.5°C. The saturation limits are at the high and low values of 10 V and 0 V. The response using the complete algorithm described here is labelled 1, the response using the correction for past saturation but not the simultaneous correction is labelled 2, and the response using only reset windup protection is labelled 3.

5. Conclusions

This paper has shown the derivation of a one step optimal correction for saturation in discrete model based controllers. The resulting controller reduces to the single input single output one step controller of Clarke and Hastings (James, 1971) and its self-tuning version (Clarke and Gawthrop, 1975), and the multi input multi-output one step controller of Koivo (1980) in the case that the controller does not saturate. The algorithm is easy to implement as a correction operating on the difference between past calculated inputs and inputs actually implemented. It is valid for any form of input saturation. A further (but usually less important) correction for multivariable controllers is developed which simultaneously adjusts the remaining inputs for any control inputs which saturate. This controller is fully equivalent to the mathematical programming solution with this objective function with the same saturation limits on the inputs. The algorithm is equally applicable to the Dahlin controller and Internal Model Controllers (IMC) because of their equivalence to a special case of the minimum variance controller. A simulation example shows the performance improvements using this algorithm for control of a multi input multi-output system.

Acknowledgments—The research work in this paper has been carried out with the support of the Natural Science and Engineering Research Council of Canada. This support is gratefully acknowledged. The authors also thank Dr C. M. Crowe for many helpful suggestions during this work.

References

- Åström, K. J. (1978). Stochastic Control Problems. In W. A. Coppel (Ed.) *A Mathematical Control Theory Proceedings*. Canberra, Australia, 1977 and in A. Dold and B. Eckmann (Eds) *Lecture Notes in Mathematics*, 680, Springer, New York.
- Bezanson, I. W. (1984). Scalar quadratic control with amplitude constraints. *Electron. Lett.*, **20**, 246–247.
- Clarke, D. W. (1981). Introduction to self-tuning controllers. In C. J. Harris and S. A. Billings (Eds), *Self-Tuning and Adaptive Controls: Theory and Applications*, in IEE Control Engineering Series 15, Peter Peregrinus, New York.
- Clarke, D. W. and P. J. Gawthrop (1975). Self-Tuning Controller. *Proc. Inst. Electron. Engng*, **122**, 929–934.
- Clarke, D. W. and R. Hastings-James (1971). Design of digital controllers for randomly disturbed systems. *Proc. Inst. Electron. Engng*, **118**, 1503–1506.
- Dahlin, E. B. (1968). Designing and choosing digital controllers. *Instrum. Control Syst.*, **4**, 77.
- Garcia, C. E. and M. Morari (1982). Internal model control. I. A unifying review and some new results. *Ind. Engng Chem. Process Des. Dev.*, **21**, 308–323.
- Gill, E., W. Murray and M. H. Wright (1981). Practical optimization. Academic Press, New York.
- Goodwin, G. C. (1972). Amplitude-constrained minimum-variance controller. *Electron. Lett.*, **8**, 187–188.
- Koivo, H. N. (1980). A multivariable self-tuning controller. *Automatica*, **16**, 352–366.
- MacGregor, J. F., T. J. Harris and J. D. Wright (1984). Duality between the control of processes subject to randomly occurring deterministic disturbances and ARIMA stochastic disturbances. *Technometrics*, **26**, 389–397.
- Mäkilä, P. M. (1982). Self-tuning control with linear input constraints. *Opt. Control App. Meth.*, **3**, 337–353.
- Parrish, J. R. and C. B. Brosilow (1985). Inferential control applications. *Automatica*, **21**, 527–538.
- Segall, N. L., J. F. MacGregor and J. D. Wright (1986). One-Step optimal correction for input saturation in discrete model-based controllers, Report #1013, Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada.
- Toivonen, H. (1983a). Suboptimal control of linear discrete stochastic systems with linear input constraints. *IEEE Trans. Aut. Control*, **AC-28**, 246–248.
- Toivonen, H. (1983b). Suboptimal control of discrete stochastic amplitude constrained systems. *Int. J. Control*, **37**, 493–502.
- Toivonen, H. (1983c). Design and analysis of discrete stochastic amplitude-constrained systems – review of procedures and computer program. University of Trondheim, Norwegian Institute of Technology, Technical report 83-47-W.
- Wong, P. M., P. A. Taylor and J. D. Wright (1987). An experimental evaluation of saturation algorithms for advanced digital controllers. *Ind. Eng. Chem. Process Des. Dev.*, **26**, 1117–1126.

Appendix A. Simultaneous correction for MIMO systems

Equation (13) in Section 2 describes the saturation correction to the controller transfer function using past unimplemented control actions U_i^j . When any inputs in a multivariable system saturate then a further simultaneous correction is required to calculate the one-step optimal settings of the control inputs subject to the saturation limits.

Consider equation (9) rearranged using equations (11) and (12) as

$$\begin{aligned} Q^{-1}B_0^{-1}J_{t,u} &= 2G(q^{-1})F^{-1}(q^{-1})Y_t + 2B(q^{-1})U_t \\ &\quad + 2C(q^{-1})F^{-1}(q^{-1})Q^{-1}B_0^{-1}R\sum^d U_t \\ &\quad + 2[I - C(q^{-1})F^{-1}(q^{-1})][B_0 + Q^{-1}B_0^{-1}R]U_t^j \end{aligned} \quad (9a)$$

where the $J_{t,u}$ is the gradient of the positive definite quadratic objective function (1). This equation can be

rewritten with obvious substitutions as

$$J_{t,u} = 2\alpha(q^{-1})Y_t + 2\beta(q^{-1})U_t + 2\gamma(q^{-1})U_t^j \quad (A1)$$

All of the past Y_t , U_t and U_t^j are known and their contribution to the gradient can be lumped together as K_t , giving

$$J_{t,u} = 2\beta_0 U_t^* + 2K_t \quad (A2)$$

When there is no saturation the optimal control inputs are calculated so as to set the gradient of the objective function to zero

$$U_t^* = -\beta_0^{-1}K_t \quad (A3)$$

Because the objective function (1) is positive definite and quadratic we know that equation (A3) defines a globally optimal input.

The saturation limits are assumed to be simple, though possibly time varying, bounds

$$U_{t,n} : U_t^* \leq U_{t,n} \quad (A4)$$

which define a convex box shape. In order to find the constrained optimum an iterative procedure for determining active constraints is necessary. In order to determine which of the constraints (A4) violated by (A3) will be active the gradient (A1) must be evaluated at the point $U_t = \text{sat}(U_t^*)$. If the component of the gradient is in the direction of the constraint then the constraint will be active if the component of the gradient is into the feasible region it will be inactive. This simple active set strategy is due to the constraints being bounds on the inputs (Gill *et al.* 1981).

If one or more constraints are active then those components of the input are fixed at their limits and the corresponding components of the $J_{t,u}$ will not be zero at the next calculated value of the input U_t . To solve for the remaining unconstrained components of U_t and the nonzero components of $J_{t,u}$ equation (A2) must be rearranged to collect these terms on the left hand side.

For example, if only the i th input variable will be saturated, we can rearrange (A2) as

$$[2\beta_0^i] \cdots [e_i] \cdots [2\beta_0^n] \begin{bmatrix} U_t^i \\ J_{t,u} \\ U_t^n \end{bmatrix} = \begin{bmatrix} 2K_t^i \\ 2K_t \\ 2K_t^n \end{bmatrix} \quad 2\beta_0^i U_t^{ii} \quad (A5)$$

where β_0^i is the i th column of β_0 and U_t^i , $J_{t,u}^i$ and K_t^i are the i th components of U_t , $J_{t,u}$ and K_t , respectively. e_i is the i th unit vector and U_t^{ii} is the saturated value of the i th component of the input. Equation (A5) can be solved for the remaining components of the input U_t and the free component of the gradient $J_{t,u}^i$. When more than one component of the input is saturated the rearrangement shown in (A5) is performed for each component. For computational purposes the inverse of the β_0 matrix can be updated using a rank one update formula for each saturated component of the input. The control inputs calculated from (A5) now become the new iterate for U_t .

The algorithm for finding the constrained optimal input is then:

- Calculate the globally optimal control input U_t^* .
- If any of the components of U_t^* violate their saturation limits set those components to their limiting values and calculate the gradient (A2) of the objective function at that point.
- For each component of the gradient which is in the direction of the constraint, that constraint is active.
- If all of the constraints are active or if no inactive constraint becomes active and if no active constraint becomes inactive then the constraint optimum has been found.
- For each active constraint set the input corresponding to that component to its saturation limit U_t^{ii} and rearrange equation (A2) as (A5).

- (vi) Solve equation (A5) for the free components of the input and the free components of the gradient.
- (vii) Go to step (ii). This algorithm converges in at most n steps.

The \mathbf{Q} and \mathbf{B}_0 matrices heavily influence the nature of the simultaneous correction. The more interactive the \mathbf{B}_0 matrix is, the larger the effect of the simultaneous correction will be.

Although the \mathbf{Q} matrix has no effect on the Minimum Variance Controller ($\mathbf{R} = 0$) of equation (18), it can have a very significant effect at saturation when the forecast of the output cannot be cancelled exactly. The weights in the \mathbf{Q} matrix determine the relative importance of the various output deviations, when calculating the simultaneous correction.

Brief Paper

Modeling of Uncertain Dynamics for Robust Controller Design in State Space*

ALTUĞ İFTAR† and UMIT ÖZGÜNER‡

Key Words—Robustness; robust control; state space methods; control system design; stability; disturbance rejection; model reduction; flexible structures

Abstract—Structured modeling of uncertain dynamics for robust controller design in state space is discussed. It is shown that, under mild conditions, it is possible to obtain a rational transfer function matrix (TFM), possibly dependent on a parameter vector which varies over a subset of a finite dimensional real vector space, to represent uncertain dynamics. Furthermore, in many practical cases, uncertain dynamics can be represented by a relatively low order TFM, even if the actual dynamics are of very high order. The procedure of determining such a TFM is discussed. It is shown that a controller which stabilizes the nominal system, including such a representation of uncertain dynamics, also stabilizes the actual system. Furthermore, desired performance or relative stability can also be guaranteed. The presented approach gives a unified framework for the solution to the robust controller design problem for systems with both parameter uncertainties and uncertain dynamics. An application to a robust controller design problem for a large flexible structure is also presented.

1. Introduction

ONE OF THE fundamental issues in feedback controller design is robustness. Since an exact model of a physical system would be very complicated if not impossible to obtain, the designer should base the controller design on a *nominal* model and should require the controller to perform satisfactorily under possible deviations from the nominal model. At this point, modeling of such deviations (i.e. *uncertainties*) becomes a crucial issue. The uncertainties in a mathematical model of a physical system are basically due to two factors: unknown values of certain system parameters (e.g. resistance of an electrical component) and totally unmodeled dynamics (e.g. high frequency modes of a flexible structure). It has been generally accepted that state space models are more suitable for representing uncertainties due to parameter variations, while unmodeled dynamics can be represented easier in the frequency domain.

Most of the frequency domain approaches introduced so far (e.g. Francis *et al.*, 1984) are designed for linear systems with truly unstructured uncertainties and cannot take advantage of all the available information. More specifically, these methods assume only a known upper bound on the magnitude of the possible perturbations. The approach introduced by Doyle (1982) can accommodate some known structure of the uncertainty by putting the system into the so-called "standard form", but again only known magnitude

bounds can be utilized on each uncertain block. However, in practice more information is generally available. For example, in the case of flexible structures with co-located actuation and sensing it is known that the phase of the uncertain dynamics always lies between 0° and 180°. Such additional information may relax the robust controller design problem considerably and would, in general, lead to less conservative controller design.

Many physical systems possess both parameters with unknown exact values and uncertain dynamics. Wei and Yedavalli (1987) proposed a combined frequency domain and state space approach for such systems. Boyd (1986) considered representing unstructured uncertainty in a structured form and hence combining both classes for a state space design. However, his approach is restricted to systems which can be transformed into the so called standard form. Furthermore, only magnitude bounds on those blocks can be utilized in the design process.

In the present paper a unified approach in state space which can be used for linear systems with both classes of uncertainties and which can utilize any information about the uncertain dynamics is introduced. The main result, presented in Section 2, demonstrates that, under mild conditions, it is possible to obtain a rational transfer function matrix (TFM), possibly parametrized by a finite dimensional vector, to represent uncertain dynamics and design a controller accordingly to satisfy stability and desired performance. The underlying idea here is to represent possibly very high dimensional unstructured uncertain dynamics in a relatively low order structured form. The presented approach is applied to a robust controller design problem for a large flexible structure in Section 3. The structure under consideration possesses both parameters with unknown exact values and unmodeled (high frequency) dynamics.

In the sequel \mathbb{C} denotes the space of complex numbers, \mathbb{C}^+ denotes the closed right half complex plane, \mathbb{R} denotes the space of real numbers, \mathbb{R}^+ denotes the set of non-negative real numbers, \mathbb{R}^k denotes the k dimensional real vector space, I denotes the identity matrix of appropriate dimensions, $[a, b]$ denotes the closed interval of \mathbb{R} from a to b , and $\angle(\cdot)$, $|\cdot|$, and $\bar{(\cdot)}$ denote, respectively, the complex conjugate, the magnitude and the phase of (\cdot) .

2. Reduced order models for uncertain dynamics

The uncertain dynamics of a linear system, nominally modeled by $G(s)$, can be represented at the input by $A_1(s)$, at the output by $A_2(s)$, or additively by $A_3(s)$ as shown in Fig. 1. For brevity, here we consider only uncertainties represented at the output. Similar results can be proved for other types of representations along the same lines. We will, in fact, apply those results to an additively modeled uncertainty in Section 3. Henceforth we drop the subscript "o" of $A_o(s)$.

Although $A(s)$ may not be known exactly, some information about it is generally available. Furthermore, $A(s)$ may not be representable as a finite dimensional linear system with a rational TFM, or such a representation may require an unmanageably high dimensional model.

Our purpose is to determine a rational TFM $T(s; p)$ to

* Received 1 March, 1988; revised 14 November, 1988; revised 22 August, 1989; received in final form 20 February, 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak.

† Department of Electrical Engineering, University of Toronto, Toronto M5S 1A4, Canada. Author to whom all correspondence should be addressed.

‡ Department of Electrical Engineering, The Ohio State University, Columbus, Ohio 43210, U.S.A.

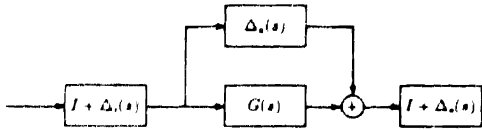


FIG. 1. Representation of uncertain dynamics.

replace $\Delta(s)$ in the model, such that a feedback controller $K(s)$ designed for such a model stabilizes the actual plant and achieves desired performance. Here $E(s; p)$ is possibly parametrized by a vector p which is taken from a parameter set Π . The following theorem demonstrates that once a suitable TFM and an associated parameter set are found to represent the uncertainties, such an approach would, in fact, guarantee the stability of the actual closed-loop system.

Theorem 1a (Absolute stability). Suppose that $\Delta(s)$ is analytic in the closed right half complex plane C^+ , except possibly at some isolated singular points which constitute a total of m poles with due count of multiplicity.* Let $E(s; p)$ be a TFM which is rational in s and continuous in p and Π be a subset of a finite dimensional real vector space with the properties:

- (a) for all $\omega \in \mathbb{R}$ there exists $p_\omega \in \Pi$ such that

$$E(j\omega, p_\omega) = \Delta(j\omega) \quad (1a)$$

if $\Delta(s)$ is analytic at $s = j\omega$ or

$$\lim_{r \rightarrow 0} E(j\omega + re^{i\theta}; p_\omega) = \lim_{r \rightarrow 0} \Delta(j\omega + re^{i\theta}) \quad \forall \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \quad (1b)$$

if $\Delta(s)$ is not analytic at $s = j\omega$.

- (b) there exists $p_\infty \in \Pi$ such that

$$\lim_{r \rightarrow \infty} E(re^{i\theta}; p_\infty) = \lim_{r \rightarrow \infty} \Delta(re^{i\theta}) \quad \forall \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right], \quad (1c)$$

and

- (c) $E(s; p)$ has exactly m poles in C^+ for all $p \in \Pi$.

For a given rational TFM $G(s)$, suppose that a rational TFM $K(s)$ is chosen such that the closed-loop TFM

$$T_r(s; p) \triangleq (I + E(s; p))G(s)K(s) \times [I + (I + E(s; p))G(s)K(s)]^{-1} \quad (2)$$

does not have any poles in C^+ for all $p \in \Pi$. Also assume that $\det[I + (I + \Delta(s))G(s)K(s)] \neq 0$ on a dense subset of C . Then the true closed-loop TFM:

$$T_A(s) \triangleq (I + \Delta(s))G(s)K(s)[I + (I + \Delta(s))G(s)K(s)]^{-1} \quad (3)$$

is analytic in C^+ .

Proof. Throughout in this proof we drop the arguments s and p for notational brevity. Let n denote the number of poles of $(I + \Delta)GK$ in C^+ . Then, since Δ and E have the same number of poles in C^+ , $(I + E)GK$ also has n poles in C^+ for all $p \in \Pi$. Furthermore, since T_r does not have any poles in C^+ , by the multivariable Nyquist stability criterion (MacFarlane, 1970), the map $\det[I + (I + E)GK]$ as s is varied on the standard Nyquist contour \mathcal{D} encircles the origin $-n$ times for all $p \in \Pi$. By the hypothesis, Δ is analytic on C^+ , except possibly at isolated singular points. The same is also true for G and K , since they are rational TFMs. Therefore, $\det[I + (I + \Delta)GK]$ is analytic on C^+ , except possibly at isolated singular points. Furthermore, since $\det[I + (I + \Delta)GK] \neq 0$ on a dense subset of C , T_A is well defined on such a set and we can apply the multivariable Nyquist stability criterion. Note that, the conditions (a) and (b) imply that the loci of $\det[I + (I + \Delta)GK]$, as s is varied on \mathcal{D} , is a subset of $\det[I + (I + E)GK]$ loci as s is varied on \mathcal{D} and p is varied on Π (see Fig. 2). Furthermore, since E is continuous in p , any member of the former set must encircle any point on the complex plane that is encircled by the latter

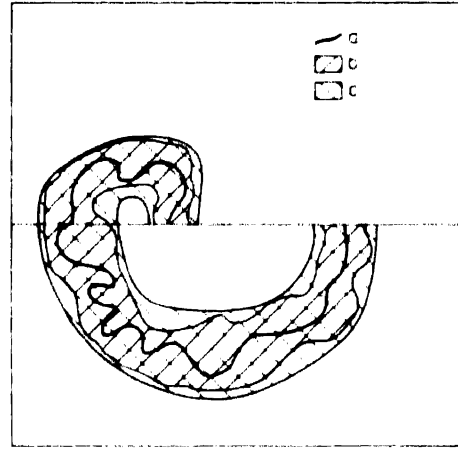


FIG. 2. (a) A representative $\det[I + (I + \Delta)GK]$ loci; (b) region of all possible $\det[I + (I + \Delta)GK]$ loci; (c) $\det[I + (I + E)GK]$ loci as s is varied on the half Nyquist contour and p is varied on Π

set the same number of times. Thus, $\det[I + (I + \Delta)GK]$ encircles the origin $-n$ times as s is varied on \mathcal{D} , which, by the multivariable Nyquist stability criterion, proves that T_A does not have any poles in C^+ . Hence, the result follows. \square

Remark 1. Note that $T_A(s)$ represents a physical system whenever $\Delta(s)$ does, since all other terms that appear in the right-hand side of (3) are rational TFMs. Furthermore, $T_A(s)$ represents the actual closed-loop system whenever $\Delta(s)$ is the TFM representing the actual unmodeled dynamics. Thus, whenever $T_A(s)$ is analytic in C^+ , the actual closed-loop system is stable.

In many practical cases, absolute stability may not be sufficient. Instead, for example, one may wish to confine all the closed-loop eigenvalues in a region \mathcal{R} of the complex plane. Note that the imaginary axis and the right-hand semi-circle of infinite radius together define the boundary of C^+ . Hence, if the conditions (a) and (b) of Theorem 1a are met, we say that $E(s; p)$ and $\Delta(s)$ are *matched* on the boundary of C^+ . Furthermore, we say that the two TFMs are *matched* on the boundary of a region \mathcal{R} , if the obvious generalizations of the conditions (a) and (b) of Theorem 1a hold. The following result can be proved as Theorem 1a, with an obvious modification of the Nyquist contour

Theorem 1b (Relative stability) Let \mathcal{R} be an open subset of C . Suppose that $\Delta(s)$ is analytic in $\mathcal{R} \triangleq C \setminus \mathcal{R}$, except possibly at some isolated singular points which constitute a total of m poles with due count of multiplicity. Let $E(s; p)$ be a TFM which is rational in s and continuous in p and Π be a subset of a finite dimensional real vector space such that $E(s; p)$ and $\Delta(s)$ are matched on the boundary of \mathcal{R} and $E(s; p)$ has exactly m poles in \mathcal{R} for all $p \in \Pi$. For a given rational TFM $G(s)$, suppose that a rational TFM $K(s)$ is chosen such that the closed-loop TFM (2) does not have any poles in \mathcal{R} for all $p \in \Pi$. Also assume that $\det[I + (I + \Delta(s))G(s)K(s)] \neq 0$ on a dense subset of C . Then the true closed-loop TFM (3) is analytic in \mathcal{R} . \square

It is well known that (Safonov *et al.*, 1981) most of the widely used performance measures (e.g. disturbance rejection and steady state error) can be related to the return difference matrix. For example, to achieve a certain degree of plant disturbance rejection at the output, one may require the return difference matrix to satisfy:

$$[I + (I + \Delta(j\omega))G(j\omega)K(j\omega)]^S \succeq Q(\omega) \quad \forall \omega \in \Omega, \quad (4)$$

provided that the left-hand side is well defined. Here $Q(\omega)$ is

* Note that, one needs to know the total number of unstable poles of $\Delta(s)$. Their actual locations need not be known.

* The more widely used criterion is defined as a lower bound on the singular values of the return difference matrix. However, note that (4) is a more general condition.

a positive definite matrix for all $\omega \in \Omega$, $\Omega \subset \mathbb{R}$ is the set of frequencies where disturbances are effective, for appropriately dimensioned matrices A and B , $A^T \triangleq A^{H^*}$, A^{H^*} denotes the complex conjugate transpose of A , and $A \succeq B$ means $A - B$ is positive semi-definite. The following theorem demonstrates that if a controller is designed to satisfy such a performance criterion for the *design model* (the model in which Δ is replaced by E), then the actual closed-loop system satisfies the same criterion

Theorem 1c (Good performance). Under the conditions of Theorem 1a, suppose $K(s)$ is chosen such that

$$[I + (I + E(j\omega; p))G(j\omega)K(j\omega)]^{-1} \succeq Q(\omega) \quad \forall \omega \in \Omega, \quad \forall p \in \Pi \quad (5)$$

Then (4) is satisfied provided that the left-hand side of (4) is well defined for all $\omega \in \Omega$.

Proof. If the left-hand side of (4) is well defined, then $\Delta(s)$ is analytic at $s = j\omega$. By (1a), for such ω there exists $p_\omega \in \Pi$ such that

$$I + (I + E(j\omega; p_\omega))G(j\omega)K(j\omega) = I + (I + \Delta(j\omega))G(j\omega)K(j\omega)$$

Hence, the result follows.

Under mild conditions on $\Delta(s)$, the existence of a TFM $E(s, p)$ and a parameter set Π , satisfying the conditions of the above theorems, can be ensured. To avoid notational complexity, we prove this only for the scalar case. The extension to the multivariable case is possible along similar lines. Furthermore, here we consider only absolute stability. A similar result can be proved for the relative stability case if the region \mathcal{R} is symmetric about the real axis.

Theorem 2. Let

$$\Delta(s) = \frac{1}{d_m(s)} \Delta_1(s) \quad (6)$$

where $d_m(s)$ is an m th order monic polynomial with zeros in C^+ and $\Delta_1(s)$ is analytic in C^+ (i.e. factor out the unstable poles of $\Delta(s)$ as $1/d_m(s)$). Furthermore, suppose that $\Delta(s)$ is such that $\Delta(j\omega) = \Delta(-j\omega)$ for all $\omega \in \mathbb{R}$, and that $\lim_{\theta \rightarrow 0} \Delta(re^{j\theta})$ is a finite real constant* for all $\theta \in [-\pi/2, \pi/2]$. Then, there exists a function $E(s, p)$, continuous in p and proper and rational in s , and a subset Π of a finite dimensional real vector space, such that conditions (a), (b) and (c) of Theorem 1a hold.

Proof. Let $p = (\alpha, \beta) = (\alpha_1, \dots, \alpha_{11}, \beta_1, \dots, \beta_m)$ and

$$E(s, p) = \frac{1}{e(s, \beta)} E_1(s, \alpha)$$

where $e(s, \beta) = s^m + \beta_1 s^{m-1} + \dots + \beta_{m-1} s + \beta_m$ and

$$E_1(s, \alpha) = \frac{\alpha_6 s^5 + \alpha_5 s^4 + \alpha_8 s^3 + \alpha_9 s^2 + \alpha_{10} s + \alpha_{11}}{s^5 + \alpha_1 s^4 + \alpha_2 s^3 + \alpha_3 s^2 + \alpha_4 s + \alpha_5}$$

For any $\omega \in \mathbb{R}^+$, let $\mathcal{H}_\omega(\omega) \subset \mathbb{C}$ be the set of all possible values of $d_m(j\omega)$. Then it is possible to find a subset $B(\omega)$ of \mathbb{R}^m , such that

- (i) the loci of $e(j\omega; \beta)$ as β is varied over $B(\omega)$ contains $\mathcal{H}_\omega(\omega)$, and
- (ii) $e(s, \beta)$ has m zeros in C^+ for all $\beta \in B(\omega)$.

For any $\omega \in \mathbb{R}^+$, let $\mathcal{A}_\omega(\omega) \subset \mathbb{C}$ be the set of all possible values of $\Delta_1(j\omega)$.† Then it is possible to find a subset $A(\omega)$ of \mathbb{R}^{11} , such that

- (i) the loci of $E_1(j\omega; \alpha)$ as α is varied over $A(\omega)$ contains $\mathcal{A}_\omega(\omega)$, and
- (ii) $E_1(s, \alpha)$ is analytic in C^+ for all $\alpha \in A(\omega)$.

Let $\mathcal{H}_\omega \subset \mathbb{R}$ be the set of all possible values of $\lim_{\theta \rightarrow 0} \Delta(re^{j\theta})$. Then it is possible to find subsets A_ω of \mathbb{R}^{11}

and B_ω of \mathbb{R}^m , such that

- (i) the loci of $\lim_{r \rightarrow \infty} E(re^{j\theta}, \alpha, \beta)$ as α is varied over A_ω and β is varied over B_ω contains \mathcal{H}_ω ,
- (ii) $E_1(s, \alpha)$ is analytic in C^+ for all $\alpha \in A_\omega$, and
- (iii) $e(s, \beta)$ has m zeros in C^+ for all $\beta \in B_\omega$.

Let

$$A \triangleq A_\omega \cup \left\{ \bigcup_{\omega \in \Omega} A(\omega) \right\}$$

$$B \triangleq B_\omega \cup \left\{ \bigcup_{\omega \in \Omega} B(\omega) \right\}$$

and

$$\Pi \triangleq A \times B$$

Then $E(s, p)$ is continuous in p and proper and rational in s , Π is a subset of a finite dimensional real vector space (\mathbb{R}^{m+11}), and the conditions (a), (b) and (c) of Theorem 1a hold. □

Remark 2. In certain cases, depending on how much is known about $\Delta(s)$, the minimum required dimension of Π (i.e. the number of parameters) can, in fact, be less than the number predicted in the above proof. On the other hand, in certain other cases, one may prefer to work with a higher order TFM and a higher dimensional parameter set, to reduce the possible conservatism involved in representing $\Delta(s)$.

Remark 3. Note that $E(s, p)$ and $\Delta(s)$ may be matched by using different $p \in \Pi$ at different points s on the boundary of C^+ (or of \mathcal{R} in the case of relative stability). However, this fact does not bring any restrictions in satisfying properties such as absolute stability, relative stability, or good performance of the actual system (see the proofs of Theorems 1a and 1c).

The following example illustrates the procedure of determining $E(s, p)$ and Π .

Example 1. Suppose all that is known about $\Delta(s)$ is that it is analytic in C^+ , $\Delta(j\omega) = \Delta(-j\omega)$, $\lim_{r \rightarrow \infty} \Delta(re^{j\theta}) = 0$ for all $\theta \in [-\pi/2, \pi/2]$

$$\left| \frac{\delta_m \omega_1}{j\omega + \omega_1} \right| \leq |\Delta(j\omega)| \leq \left| \frac{\delta_M \omega_1}{j\omega + \omega_1} \right| \quad \forall \omega \in \mathbb{R}^+,$$

and

$$\angle \left(\frac{\delta_m \omega_1}{j\omega + \omega_1} \right) \leq \angle(\Delta(j\omega)) \leq \angle \left(\frac{\delta_M \omega_1}{j\omega + \omega_1} \right) \quad \forall \omega \in \mathbb{R}^+,$$

where $\delta_M = \delta_m > 0$ and $\omega_1 = \omega_1 > 0$ are known numbers. The gain and phase of all possible $\Delta(s)$, together with a representative $\Delta(s)$ are depicted in Fig. 3. Let us choose

$$E(s, p) \quad (7)$$

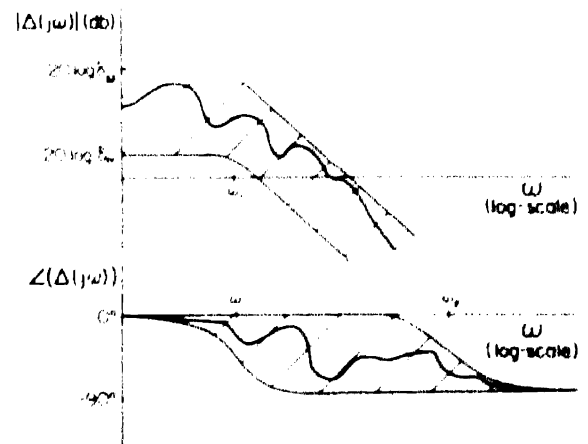


Fig. 3 Regions of all possible magnitude and phase of $\Delta(j\omega)$ and a representative $\Delta(j\omega)$ for Example 1.

* The actual value of this constant need not be known.

† $\mathcal{A}_\omega(\omega)$ is well defined since $\Delta_1(s)$ is analytic on the imaginary axis $j\mathbb{R} \subset C^+$.

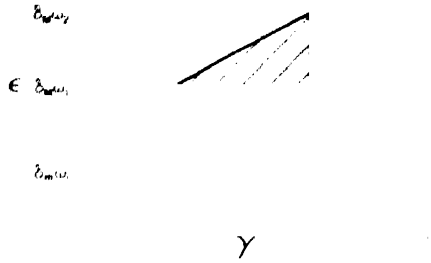


FIG. 4 The parameter set Π for Example 1

where $p = (\gamma, \epsilon)$. For each frequency $\omega \in \mathbf{R}^+$,

- by varying γ within the interval $[\omega_1, \omega_2]$ we can satisfy the phase condition $\angle(E(j\omega, p)) = \angle(\Delta(j\omega))$ and
- for a fixed $\gamma \in [\omega_1, \omega_2]$, we can meet the magnitude condition $|E(j\omega, p)| = |\Delta(j\omega)|$ by varying ϵ within the interval $[\delta_m \omega_1, \delta_M \gamma]$

for all possible $\Delta(s)$. Hence, we obtain

$$\Pi = \{(\gamma, \epsilon) \mid \omega_1 \leq \gamma \leq \omega_2, \delta_m \omega_1 \leq \epsilon \leq \delta_M \gamma\} \quad (8)$$

which is a bounded subset of \mathbf{R}^2 and is depicted in Fig. 4. \square

Remark 4 Note that, the tightest H^∞ norm bound for $\Delta(s)$ described in the above example is δ_M . Suppose that a controller was to be designed, say to achieve robust stability, for a plant with uncertainty $\Delta(s)$ by using the H^∞ approach (Francis *et al.*, 1984). Then it would be necessary to design a controller to stabilize all the plants with uncertainty $E(s)$ satisfying

$$\|E(s)\|_\infty = \delta_M \quad (9)$$

where $\|\cdot\|_\infty$ denotes the usual H^∞ norm. This, however, is a larger class than the one described by (7) and (8). Therefore, a controller designed by using the H^∞ approach would, in general, be more conservative than a controller designed using the present approach. It is, of course, possible to modify the nominal TFM and reduce the bound in (9); however, it is apparent that by using the norm bounds alone (as in the H^∞ approach) one cannot get a tighter representation (of the actual uncertainty) than the representation obtained by the present approach.

Consider a system modeled by a nominal TFM $G(s, p_c)$, where $p_c \in \Pi_c$ is a vector denoting possible values of some system parameters. Suppose that the dynamics not modeled by G can be represented at the output by an uncertain TFM $\Delta(s)$ satisfying the conditions of Theorem 2. Then, one can obtain a rational TFM $E(s, p_f)$ and a corresponding parameter set Π_f , satisfying the conditions of Theorem 1a (or of Theorem 1b when appropriate), as discussed above. Thus, a design model described by the TFM $(I + E(s; p_f))G(s; p_c)$ is obtained. A state space realization, which is parametrized by $p \triangleq (p_f, p_c) \in \Pi \triangleq \Pi_f \times \Pi_c$, can now be developed for this model. Once such a model is developed, existing methods (e.g. Karamarkar and Šiljak, 1979; Ackermann, 1980; Yedavalli, 1986; Keel *et al.*, 1988) can be used to design a controller $K(s)$ to achieve robust stability and/or good performance for all $p \in \Pi$. Then the actual closed-loop system, shown in Fig. 5, has the desired robust stability and/or good performance properties.

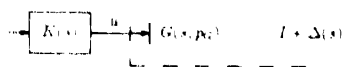


FIG. 5 Closed-loop system

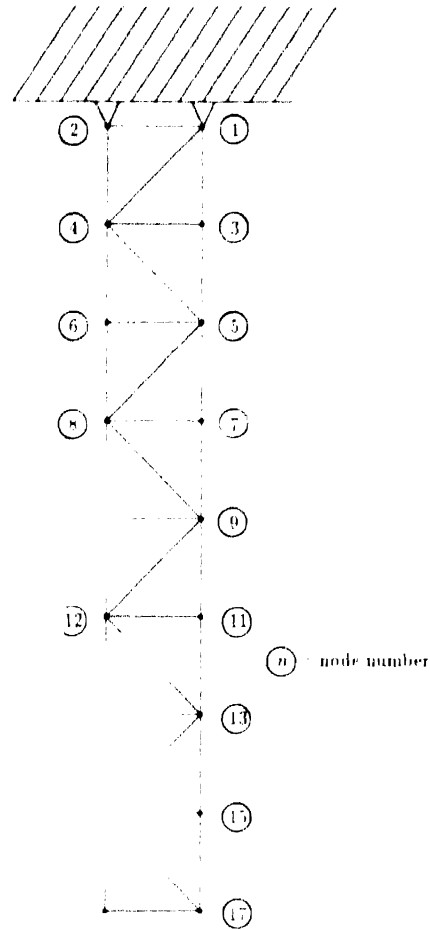


FIG. 6 Truss structure

3 Application

We consider the application of the presented design approach to the planar truss structure shown in Fig. 6. The truss is made of identical uniform aluminum rods that can be displaced in the axial direction. The structure is fixed at nodes 1 and 2. It is controlled by a linear force actuator located at node 16, acting in the horizontal direction. The location of the actuator is chosen such that it has the greatest effect on the first mode relative to the higher modes. Measurements are taken by a co-located linear accelerometer. Ideal sensor and actuator dynamics are assumed. The nominal values of the relevant material parameters for the rods are given in Table 1.

The structure has 16 free nodes, each having two degrees of freedom. Hence, it exhibits 32 flexible modes. It is assumed that a certain structural damping is associated with each individual mode. These assumptions lead to a 64th order state space model. Due to variations of material properties (such as modulus of elasticity and mass density) the exact values of structural frequencies ω_i and mode shapes b_i are uncertain. Nominal values of these quantities (based on the values given in Table 1) for the selected modes are given in Table 2. The actual values are assumed to be within $\pm 1\%$ of these nominal values. The damping ratios ζ_i ($i = 1, 2, \dots, 32$) are assumed to be between 0.005 and 0.01.

TABLE 1. MATERIAL PARAMETERS

Rod cross-section area:	$4 \times 10^{-4} \text{ m}^2$
Modulus of elasticity:	$6.8944 \times 10^{10} \text{ N/m}^2$
Mass density:	2750 kg/m^3
Length of a vertical or horizontal rod:	1 m

TABLE 2. MODAL FREQUENCIES AND MODAL SHAPES

i	ω_i (rad/sec)	b_i
1	131	0.3927
2	633	0.1403
3	843	0.0223
...
31	15923	-0.1731
32	16602	-0.3115

It is desired to design a controller to actively dampen the first mode while maintaining total system stability. The modeled dynamics are assumed to be associated with the first mode only. The uncertainties in ω_1 , b_1 and ζ_1 are treated here as parameter uncertainties and the part of the system associated with the higher modes is treated as unmodeled dynamics. The transfer function description of this model is

$$G_m(s; p_G) = \frac{b_1^2 s^2}{s^2 + 2\zeta_1 \omega_1 s + \omega_1^2}, \quad p_G = (\omega_1, b_1, \zeta_1) \in \Pi_1, \tag{10}$$

where

$$\Pi_1 = \{(\omega_1, b_1, \zeta_1) \mid \omega_1 \in [\omega_{1min}, \omega_{1max}], \zeta_1 \in [\zeta_{1min}, \zeta_{1max}], b_1 \in [b_{1min}, b_{1max}]\} \tag{11}$$

The bounds on the individual parameters are readily obtained utilizing the previous assumptions (see Ilar, 1988 for details).

The transfer function of the actual system is $G(s) = \sum_{i=1}^N G_i(s)$, where

$$G_i(s) = \frac{b_i^2 s^2}{s^2 + 2\zeta_i \omega_i s + \omega_i^2}, \quad i = 1, 2, \dots, 32 \tag{12}$$

Therefore if we let $G(s) = G_m(s) + \Delta(s)$, then we obtain a description of unmodeled additive dynamics as $\Delta(s) = \sum_{i=2}^N G_i(s)$. The magnitude and phase of $\Delta(j\omega)$, for typical ω_1 , b_1 and ζ_1 are depicted in Fig. 7. Although the exact plots depend on specific parameter values, the general shape of these plots are the same for all possible ω_1 , b_1 and ζ_1 values. We choose

$$E(s; p_E) = \frac{K_E s^2}{s^2 + 2\zeta_E \omega_E s + \omega_E^2}, \quad p_E = (K_E, \omega_E, \zeta_E) \in \Pi_E \tag{13}$$

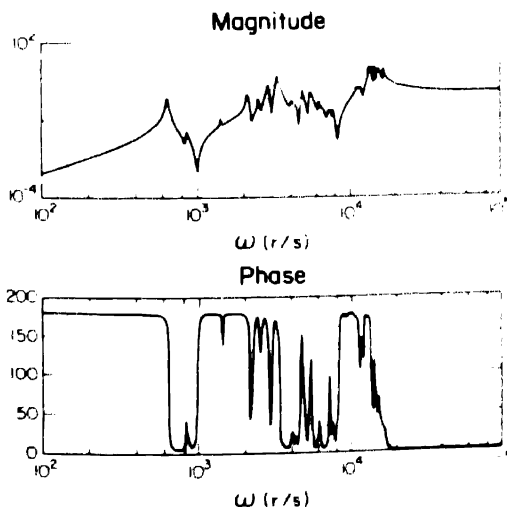


FIG. 7. Magnitude and phase of uncertain dynamics.

to represent the unmodeled dynamics, where

$$\Pi_E = \{(K_E, \omega_E, \zeta_E) \mid 0 < K_E \leq K_{Emax}, \omega_{Emin} \leq \omega_E \leq \omega_{Emax}, \zeta_{Emin} \leq \zeta_E \leq \zeta_{Emax}\} \tag{14}$$

$\omega_{Emin} = 0.99\omega_{1max} = 627$ and $\omega_{Emax} = 1.01\omega_{1max} = 16768$. An upper bound K_{Eu} for K_E can be calculated as

$$\left\{ b_i^2 + 2\zeta_i \omega_i \sum_{j=1}^N \frac{b_j^2 \omega_j^2}{\omega_j^2 - \omega_i^2} \right\} + 0.327 \tag{15}$$

Although less conservative bounds on K_E can be generated at the expense of more involved calculations, we found these bounds satisfactory for our purposes. We note that the transfer function (13) together with the set Π_E given in (14) satisfies the conditions of Theorem 1a.

We obtain a design model as a realization of $G_d(s; p_d) = F(s; p_d)$. This model depends on the parameter vector $p_d = (p_1, p_2)$ which can vary over the set $\Pi_d \triangleq \Pi_1 \times \Pi_E$. A first order controller

$$K(s) \tag{16b}$$

is designed based on this model. The controller parameters are chosen to be $a = 30$ and $K = 200$, to satisfy the design goals. The closed loop eigenvalues for the nominal system are $\{-31.72, -15.58 \pm 126.35i, -40.69 \pm 3242i\}$. It is observed that more than 12% damping has been achieved in the first mode.

To prove the stability of the actual closed loop system, it suffices to show that the closed loop design model is stable for all possible parameter variations. To accomplish that, we consider the characteristic polynomial of this model and form the four Kharitonov polynomials (Kharitonov, 1978) as a function of ω_1 . By applying Routh's stability criterion, we show that each of these four polynomials is stable for all $\omega_1 \in [622, 16768]$ (see Ilar, 1988 for details). Hence, since $\omega_{Emin} = 622$, by Kharitonov's Theorem (Kharitonov, 1978), we conclude that the closed loop design model is stable for all parameter values in Π . This guarantees the stability of the actual controlled system (by the extension of Theorem 1a to the case of additive uncertainty).

The eigenvalues of the closed-loop system with the controller (16a) applied to the 64th order "truth" model are also calculated for verification purposes. Selected eigenvalues are $\{-31.73, -15.56 \pm 126.35i, -6.703 \pm 633.22i, -122.4 \pm 15922i, -134.2 \pm 16601i\}$. It is observed that the desired damping in the first mode and overall stability are both achieved.

4. Conclusion

Modeling of uncertain dynamics for robust controller design in state space has been discussed. It has been shown that, under mild conditions on uncertain dynamics, it is possible to obtain a rational IFM, possibly dependent on a parameter vector which varies over a subset of a finite dimensional real vector space, to represent uncertain dynamics. Furthermore, in many practical cases, uncertain dynamics can be represented by a relatively low order IFM, even if the actual dynamics are of very high order. The procedure of determining such a IFM has been discussed. It has been shown that a controller which stabilizes the nominal system, including such a representation of uncertain dynamics, also stabilizes the actual system. Furthermore, desired performance or relative stability can also be guaranteed.

Once a parameter dependent rational IFM is obtained to describe the uncertain dynamics, a parameterized state space design model for the overall system can be obtained. Then, already existing methods can be used to design robust controllers. The presented approach gives a unified framework for the solution to the robust controller design problem for systems with both parameter uncertainties and uncertain dynamics.

Acknowledgements—The authors are grateful to Professor Enrique Barbieri for his generous help in modeling the truss structure of Section 3, to Professor Levent Acar for his valuable comments, and to Professor Malcolm C. Smith for helpful discussions.

References

- Ackermann, J. (1980). Parameter space design of robust control systems. *IEEE Trans. Aut. Control*, **AC-25**, 1058–1072.
- Boyd, S. (1986). A note on parametric and nonparametric uncertainties in control systems. *Proc. Am. Control Conf.*, pp. 1847–1849, Seattle, WA.
- Doyle, J. C. (1982). Analysis of feedback systems with structured uncertainties. *Proc. IEE, Part D*, **129**, 242–250.
- Francis, B. A., J. W. Helton and G. Zames (1984). \mathcal{H}^∞ -Optimal feedback controllers for linear multivariable systems. *IEEE Trans. Aut. Control*, **AC-29**, 888–900.
- Ifar, A. (1988). Robust controller design for large scale systems. Ph.D. Dissertation, The Ohio State University, Columbus, OH.
- Karmarkar, J. S. and D. D. Siljak (1979). A computer aided design of robust regulators. *Proc. IFAC Workshop*, pp. 49–58, Denver, CO.
- Keel, L. H., S. P. Bhattacharyya and J. W. Howze (1988). Robust control with structured perturbations. *IEEE Trans. Aut. Control*, **AC-33**, 68–78.
- Khantonov, V. L. (1978). Asyptotic stability of an equilibrium position of a family of systems of linear differential equations. *Differentsialnina Uravnenia*, **14**, 2086–2088.
- MacFarlane, A. G. J. (1970). Return-difference and return-ratio matrices and their use in analysis and design of multivariable feedback control systems. *Proc. IEE*, **117**, 2037–2049.
- Safonov, M. G., A. J. Laub and G. L. Hartmann (1981). Feedback properties of multivariable systems: the role and use of return difference matrix. *IEEE Trans. Aut. Control*, **AC-26**, 47–65.
- Wei, K. H. and R. K. Yedavalli (1987). Robust stabilizability for systems with both parameter variation and unstructured uncertainty. *Proc. IEEE Conf. on Decision and Control*, pp. 2082–2087, Los Angeles, CA.
- Yedavalli, R. K. (1986). Dynamic compensator design for robust stability of linear uncertain systems. *Proc. IEEE Conf. on Decision and Control* pp. 34–36, Athens, Greece.

Brief Paper

Robust Absolute Stability of Lur'e Control Systems in Parameter Space*

A. TESI† and A. VICINO‡

Key Words—Absolute stability, Lur'e control systems, Popov criterion, robust stability, parameter variations, uncertainty sets, positive realness

Abstract—This paper deals with the problem of robust absolute stability analysis for nonlinear Lur'e control systems in the presence of system parameter variations. The well known Popov criterion for absolute stability is used in order to characterize the boundary of the region of absolute stability in the parameter plane when the coefficients of the transfer function of the linear plant are polynomial functions of the uncertain parameters. For a scalar parameter, a method is given to determine the maximal interval of variation around a fixed nominal value preserving absolute stability. This result is also used to derive a technique for checking absolute stability of Lur'e systems with parameters in given planar uncertainty sets. Numerical examples showing the application of the method are reported.

1 Introduction

THE REINFORCED interest in the problem of linear systems stability analysis in parameter space in last years has brought into fashion much of the work done since several decades ago in linear and nonlinear systems robust stability analysis [see e.g. Neimark (1949) and in particular Siljak (1969, 1989a) for extensive lists of references on this subject]. While a great number of contributions have appeared recently on robust stability analysis of linear systems against parametric perturbations, there has not been a similar explosion in the area of nonlinear systems analysis and control. One of the main reasons is certainly the fact that while for linear systems necessary and sufficient conditions for asymptotic stability are well known and relatively simple, on the whole, in the nonlinear case usually only sufficient conditions for asymptotic stability are known. A practical implication of this fact is that in general it is not possible to characterize the exact region of absolute stability in parameter space. However, relatively less complicated is the description of subsets of the domain of absolute stability. In this case it is recognized as a widely open problem to ascertain to which extent these subsets approximate the true absolute stability domain, i.e. how far the available absolute stability sufficient conditions are from being also necessary.

An important class of nonlinear control systems is that of Lur'e-Postnikov systems [see e.g. Siljak (1969)]. For these systems, several sufficient conditions for absolute stability have been given since the beginning of the sixties [see e.g. Narendra and Taylor (1973); also very recent contributions can be found in the literature, e.g. Voronov (1989)]. The most widely used sufficient condition is undoubtedly that

stated by the celebrated Popov criterion (Popov, 1962). This condition reduces the absolute stability problem to positive realness of a suitable function of frequency. Few contributions to the study of positive realness in the presence of parametric uncertainty can be found in the recent literature. Bose and Delansky (1989) give easy conditions to check positive realness of a rational function whose numerator and denominator polynomials are interval polynomials. In Siliak (1989b) it is shown how convexity of the domain of positivity in polynomial coefficient space allows one to check positivity of a polytope of polynomials by considering only its vertices. In particular, in this work implications of this result on the study of robust absolute stability of certain classes of perturbed Lur'e systems are discussed.

In this paper we consider Lur'e type control systems in which either the linear part is affected by parametric uncertainties or the feedback nonlinearity sector is unknown. For these cases, we give an analytical method to study the robustness of absolute stability in the face of perturbations, when one or two physical parameters causes uncertainty in the control system. More precisely, we solve the following problems:

- Consider a Lur'e system with a feedback nonlinearity in a prescribed sector and with the linear part transfer function coefficients (or state equation matrix entries) depending on a scalar physical uncertain parameter according to polynomial functions. Given a nominal value of the parameter for which the Popov criterion is satisfied (and hence the system is absolutely stable), compute the maximum interval around this value for which the Popov criterion is satisfied.

The solution of this problem provides a region of absolute stability which in general is not the true one, due to the inherent conservativeness of the Popov criterion. A case of special interest arises when the nonlinearity sector represents the uncertain parameter. For this problem, which can be given a graphical solution in terms of the well known Popov locus of the linear part, it is shown how the proposed general analytical procedure simplifies and how it can be interpreted in terms of the Popov locus in the complex plane.

- Given a Lur'e system as in the preceding point, with the only difference being that two physical uncertain parameters and a planar uncertainty set are involved, check if the Popov criterion is satisfied for all parameter values in the uncertain set.

The paper is organized as follows. Section 2 presents the problem formulation and basic results necessary for the successive development. Sections 3 and 4 solve the problems addressed in the two points above respectively, while Section 5 reports two numerical examples showing the application of the proposed method.

2 Problem formulation and basic results

We consider Lur'e type control systems given by

$$\begin{cases} \dot{x} = Ax + bu \\ y = c^T x \\ u = -f(y) \end{cases} \quad (1)$$

* Received 31 July 1989; revised 7 February 1990; received in final form 9 April 1990. The preliminary version of this paper was presented at the 11th IFAC World Congress, Tallinn, Estonia in 1990. This paper was recommended for publication in revised form by Associate Editor P. Dorato under the direction of Editor H. Kwakernaak.

† Dipartimento di Sistemi e Informatica, Università di Firenze, Via di Santa Marta, 3-50139 Firenze, Italy.

‡ Author to whom all correspondence should be addressed: Prof. A. Vicino, Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, $b, c \in \mathbb{R}^n$, $f(\cdot): \mathbb{R} \rightarrow \mathbb{R}$ is a continuous function belonging to the class \mathcal{F}_k defined as follows

$$\mathcal{F}_k = \{f(\cdot): f(0) = 0, 0 \leq yf(y) \leq ky^2, 0 < k < \infty\} \tag{2}$$

Let us introduce the transfer function $G(s)$

$$G(s) = c^T (sI - A)^{-1} b = \frac{N(s)}{D(s)} \tag{3}$$

which represents the input-output realization of the linear plant in system (1). We make the assumption that the linear plant (A, b, c) is controllable, observable and asymptotically stable. The Lur'e control system (1) is said to be *Absolutely Stable* (AS) if the equilibrium state $x = 0$ is asymptotically globally stable, for each $f(\cdot) \in \mathcal{F}_k$.

It is well known that the fundamental Popov's theorem (Popov, 1962) provides one of the most widely used *sufficient conditions* for (1) to be AS in terms of the sector bound k and the transfer function $G(s)$. The following theorem states the Popov criterion in terms of a polynomial function.

Theorem. If there exists a real $\theta \in \mathbb{R}$ such that the following inequality holds

$$P(\omega^2) = |D(j\omega)|^2 \{k^{-1} + \operatorname{Re}\{(1 + j\omega\theta)G(j\omega)\}\} > 0, \quad \forall \omega \geq 0 \tag{4}$$

then system (1) is absolutely stable. □

The above theorem relates AS to *positivity* of the polynomial $P(\omega^2)$, which will be called *Popov polynomial*. We recall that a polynomial $f(x)$ is said to be positive if

$$f(x) > 0, \quad \forall x \in \mathbb{R}^+$$

It is clear that the study of robust absolute stability against possible variations of parameters in control systems (1) is equivalent to studying robust positivity of families of polynomials with perturbed coefficients. We denote by $p \in \mathbb{R}^q$ a parameter vector and by $P(\omega^2; p)$ the corresponding Popov polynomial with coefficients depending on p . Several variables can be interpreted as parameters in $P(\omega^2; p)$. Typical components of vector p may be physical uncertain parameters entering the transfer function coefficients (an uncertain transfer function will be denoted by $G(s; p)$) or the linear plant state matrices, the sector bound k or the Popov parameter θ . Of course, for a fixed sector bound k , the Popov condition (4) can be easily extended to include parameter dependence as follows

For a given uncertainty set U_p in parameter space and a fixed sector bound k , the Lur'e system is AS if the following conditions are verified $\forall p \in U_p$.

$$G(s; p) \text{ is asymptotically stable} \tag{5}$$

$$\exists \theta \in \mathbb{R}: P(\omega^2; p) > 0, \quad \forall \omega \geq 0. \tag{6}$$

We consider the case in which the coefficients of $G(s; p)$ depend *polynomially* on the parameters p . Under this assumption, the Popov polynomial to be tested for positivity can be expressed as

$$P(\omega^2; p) = \sum_{i=0}^n a_i(p) \omega^{2i} \tag{7}$$

where coefficients $a_i(p)$, $i = 0, 1, \dots, n$ are polynomial functions in p .

We will assume that condition (5) is satisfied. In fact, this assumption can be checked for several classes of perturbations (see e.g. Bhattacharyya 1987; Tesi and Vicino, 1988; Vicino, 1989) for the case in which parameters enter linearly the coefficients of $G(s; p)$ and Sideris and Peña, 1988; Genesio and Tesi, 1988; Vicino *et al.*, 1988 for the more complicated case where these coefficients are polynomial or rational functions of p).

The domain of *positivity* D_+ , whose boundary is denoted by ∂D_+ , plays a key role

$$D_+ = \{p \in \mathbb{R}^q: P(\omega^2; p) > 0, \forall \omega \geq 0\}. \tag{8}$$

Unfortunately, it is rather difficult to transform the above definition in an analytical description easily usable for testing positivity of a given uncertainty set U_p . However, if coefficients $a_i(p)$ are linear in p , then D_+ is *convex*. In fact, it follows from (8) that if p' and $p'' \in D_+$, then $(1 - \lambda)p' + \lambda p'' \in D_+$, $\forall \lambda \in [0, 1]$. This consideration allows to conclude that positivity, and hence AS, of polytopes in parameter space is implied by positivity of the polytope vertices. This important property for a polynomial can be verified by a Routh-like test (Šiljak, 1971). Unfortunately, the linearity hypothesis restricts the applicability of the above result only to very special classes of problems, where parameters p_i enter linearly the numerator coefficients of $G(s; p)$.

To investigate the polynomial dependence case, we use an analytical description of sets including ∂D_+ . Consider the envelope $E(p)$ of the family of surfaces in parameter space generated by the equation $P(\omega^2; p) = 0$ for $\omega \geq 0$. Such an envelope, if it exists, is defined by the one parameter family of real solutions of the equation system (Šiljak, 1969)

$$\begin{cases} P(\omega^2; p) = 0 \\ \partial P(\omega^2; p) / \partial \omega = 0 \end{cases}, \quad \omega \geq 0 \tag{9}$$

where $\omega \geq 0$ is the parameter.

Excluding the special case $\omega = 0$ (which can be accounted for separately) and setting $\Omega = \omega^2$, (9) can be written as

$$\begin{aligned} F_1(\Omega; p) &= \sum_{i=0}^n a_i(p) \Omega^i = 0 \\ F_2(\Omega; p) &= \sum_{i=0}^{n-1} (i+1) a_{i+1}(p) \Omega^i = 0 \end{aligned}, \quad \Omega \tag{10}$$

The first step we take now is to give an implicit representation of the envelope $E(p)$. This can be done by using a well known theorem on the resultant of two polynomials in an independent variable with indeterminate coefficients (see e.g. Jacobson, 1964). By considering the two polynomials in (10), we define the *resultant* of F_1 and F_2 with respect to the indeterminate Ω as follows

$$R_\Omega(p) = \det [H(p)] \tag{11}$$

where the matrix $H(p) \in \mathbb{R}^{2n \times 2n}$ is given by

$$H(p) = \begin{bmatrix} a_n(p) & a_{n-1}(p) & a_{n-2}(p) & a_{n-3}(p) & \dots & a_1(p) & a_0(p) \\ 0 & a_n(p) & a_{n-1}(p) & a_{n-2}(p) & \dots & a_1(p) & a_0(p) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & a_{n-1}(p) & a_{n-2}(p) & a_{n-3}(p) & a_{n-4}(p) & a_{n-5}(p) \\ na_n(p) & (n-1)a_{n-1}(p) & a_{n-2}(p) & 0 & 0 & 0 & 0 \\ 0 & na_n(p) & 2a_{n-1}(p) & a_{n-2}(p) & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & na_n(p) & (n-1)a_{n-1}(p) & 0 & a_{n-2}(p) & 0 \end{bmatrix} \tag{12}$$

The following theorem allows one to solve system (10)

Theorem [see (Jacobson, 1964) for a general statement]. $R_\Omega(p) = 0$ if and only if one of the conditions occurs:

- $a_n(p) = 0$
- The two polynomials have a common root in a suitable extension field. □

As a consequence of this theorem, the envelope $E(p)$ can be described as

$$E(p) = E_R(p) \cup E_0(p) \cup E_n(p) \tag{13}$$

where

$$\begin{aligned} E_R(p) &= \{p \in \mathbb{R}^q: R_\Omega(p) = 0 \\ &\text{and } \exists \Omega > 0: F_1(\Omega; p) = 0 \text{ and } F_2(\Omega; p) = 0\} \end{aligned} \tag{14}$$

$$E_0(p) = \{p \in \mathbb{R}^q: a_0(p) = 0\} \tag{15}$$

$$E_n(p) = \{p \in \mathbb{R}^q: a_n(p) = 0\}. \tag{16}$$

Observation 1. Observe that $p \in E_R(p)$ if and only if there

exists $\Omega > 0$ which solves simultaneously both equations in (10). Since the second equation in (10) is the derivative of the first one with respect to Ω , it follows that corresponding to $p \in E_R(p)$, the polynomial $F_1(\Omega; p)$ must necessarily have a zero of multiplicity at least 2. Moreover, if $F_1(\Omega; p)$ has, for a fixed p , r positive real zeros ($\Omega_1, \dots, \Omega_r$) of multiplicity (μ_1, \dots, μ_r) ($\mu_i \geq 2$) respectively, then the resultant $R_\Omega(p)$ has at p a zero of multiplicity $\sum_{i=1}^r \mu_i - r$.

3 Robust stability against scalar perturbations

In this section we give a solution to the following problem. Let k be a fixed sector bound and $p = [p_1, p_2]^T$ a vector with p_1 representing the Popov parameter θ and p_2 a "physical" uncertain parameter entering polynomially in system matrix A and/or vectors b, c . Let a nominal value of the parameter $p_2 = p_2^0$ be given (corresponding to the nominal linear plant) such that the Popov inequality (4) is satisfied.

With reference to the perturbed Popov polynomial $P(\omega^2; p)$, we want to evaluate the maximal connected domain of positivity of the variable p_2 , containing the nominal parameter value p_2^0 . To do this, we consider equations (13)–(16) defining the envelope containing ∂D . First of all, observe that the Popov function is linear in p_1 and polynomial in p_2 , so that $P(\omega^2; p)$ can be written as follows

$$P(\omega^2; p) = f_1(\omega^2; p_2)p_1 + f_0(\omega^2; p_2) \quad (17)$$

where $f_0(\cdot)$ and $f_1(\cdot)$ are suitable polynomial functions. We select the extremal values of p_2 considering points in the plane belonging to certain sets, called "critical" sets. For simplicity of notation, only the second components of the corresponding points will be included in these sets, denoted by $S_i(p_2)$, $i = 1, 2, 3$. The first critical set $S_1(p_2)$ is obtained by collecting points in the plane (p_1, p_2) belonging to $E(p)$ and satisfying necessary conditions for the existence of a horizontal line (parallel to the axis p_1) tangent to $E(p)$, i.e. such that

$$\begin{cases} P(\omega^2; p) = 0 \\ \partial P(\omega^2; p) / \partial p_1 = 0 \\ \partial P(\omega^2; p) / \partial \omega = 0 \end{cases} \quad (18)$$

By substituting (17) in (18), the elements p_2^i belonging to the first critical set $S_1(p_2)$ can be obtained by solving the following polynomial equation system

$$\begin{cases} f_1(\omega^2; p_2) = 0 \\ f_0(\omega^2; p_2) = 0 \end{cases} \quad (19)$$

Observe that, once the above system has been solved for p_2 and ω , the last equation in (18) provides, if desired, the corresponding values of p_1 . System (19) can be solved by using the resultant theorem reported in Section 2. In particular, in this case the resultant of the system with respect to the indeterminate ω^2 depends only on p_2 , so that we need to find real solutions of the equation

$$R_{\omega^2}(p_2) = \det[H(p_2)] = 0 \quad (20)$$

and choose only those which solve simultaneously both equations (19) for some positive value ω^2 . Notice that $H(\cdot)$ is defined as in (12). Of course, the two special cases $\omega = 0$ and $\omega = \infty$ must be considered separately, since in both cases degeneracies of the equation system (19) occur.

The second critical set $S_2(p_2)$ is obtained by computing critical points of $R_\Omega(p_1, p_2) = 0$, i.e. self intersection points of the curve of $R_\Omega(p_1, p_2) = 0$. These points are obtained by imposing that the gradient of $R_\Omega(p_1, p_2)$ is null

$$\begin{cases} \partial R_\Omega(p_1, p_2) / \partial p_1 = 0 \\ \partial R_\Omega(p_1, p_2) / \partial p_2 = 0 \end{cases} \quad (21)$$

Again, the resultant theorem allows one to solve the above system. The critical set $S_2(p_2)$ includes all solutions p_2^i of (21), which solve also (10) for positive Ω .

Observation 2. The above conditions for the existence of solutions in $S_2(p_2)$ become of easy interpretation if the order

of the linear plant is $n = 2$ (or $n = 3$). In fact, from Observation 1 of Section 2 it follows that for $n = 2$, $S_2(p_2)$ must necessarily be empty; for $n = 3$, $S_2(p_2)$ is nonempty if and only if there exists a p such that the polynomial equation $F_1(\Omega; p) = 0$ in the indeterminate Ω has one only positive root of multiplicity 3.

The last critical set $S_3(p_2)$ is obtained by looking for real solutions of the following two equation systems (compare (13, 15, 16))

$$\begin{cases} R_\Omega(p_1, p_2) = 0 \\ a_\Omega(p) = 0 \end{cases} \quad (22)$$

$$\begin{cases} R_\Omega(p_1, p_2) = 0 \\ a_\Omega(p) = 0 \end{cases} \quad (23)$$

and selecting as elements of $S_3(p_2)$ only those solutions which solve both equations (10) simultaneously for positive values of Ω .

Define now the following sets

$$S(p_2) = S_1(p_2) \cup S_2(p_2) \cup S_3(p_2) \quad (24)$$

Let us now introduce the definition of an "extremal" point of the parameter plane. We say that a point $p = (p_1, p_2)$ of the plane such that $p \in S(p_2)$ is "extremal" if it satisfies the two properties Π_1 and Π_2

$$\Pi_1 \quad p \in \partial D,$$

i.e. $P(\omega^2; p)$ is nonnegative

$$\Pi_2$$

$$\forall \epsilon > 0 \text{ arbitrarily small } p + \epsilon(p_1 - p_1^*, p_2 - p_2^*) \notin D,$$

$$\text{and } p - \epsilon(p_1 - p_1^*, p_2 - p_2^*) \notin D,$$

i.e. $P(\omega^2; p + \epsilon)$ and $P(\omega^2; p - \epsilon)$ are not positive

Observe that these properties can be easily checked by means of algebraic nonnegativity tests (Siljak, 1971).

Let us define the "extremal" set $S_e(p_2)$ as follows

$$S_e(p_2) = \{p \in S(p_2) \mid (p_1, p_2) \text{ is extremal}\}$$

and the two subsets $S_e^+(p_2)$ and $S_e^-(p_2)$ as

$$S_e^+(p_2) = \{p \in S_e(p_2) \mid p_2 \geq p_2^0\} \quad (25)$$

$$S_e^-(p_2) = \{p \in S_e(p_2) \mid p_2 \leq p_2^0\} \quad (26)$$

and the following quantities

$$p_2^m = \max_{p \in S_e^+(p_2)} p_2^i \quad (27)$$

$$p_2^M = \min_{p \in S_e^-(p_2)} p_2^i \quad (28)$$

In the above equations $p_2^M(p_2^m)$ is set to $+\infty$ ($-\infty$) if the set $S_e^-(p_2)$ ($S_e^+(p_2)$) is empty. Assuming that $G(s; p)$ is asymptotically stable for $p \in (p_2^m, p_2^M)$, the following theorem follows readily from the previous considerations.

Theorem. The parameter uncertainty domain (p_2^m, p_2^M) is the maximal connected domain of absolute stability, according to the Popov criterion, containing the nominal parameter for a given Lur'e control system. \square

Notice that maximality of the domain of absolute stability computed above does not necessarily mean that it is not possible to extend the interval (p_2^m, p_2^M) still preserving AS, because the Popov condition is sufficient but not necessary for AS.

3.1 Maximal sector of absolute stability. In this subsection, we briefly show how the procedure given before simplifies when we consider the special case in which the parameter p_2 represents the reciprocal of the sector bound k . In this case, we assume for coherence with (2) that the allowed minimum k is 0 and accordingly $p_2^m = \infty$. We want to estimate the maximal k for which AS is preserved according to the Popov criterion.

The Popov polynomial turns out to be linear in both parameters $p_1 = \theta$ and $p_2 = k^{-1}$. The envelope equations

(10) become

$$\begin{cases} f_1(\omega^2)p_1 + f_2(\omega^2)p_2 + f_0(\omega^2) = 0 \\ (1/\omega)[f_1'(\omega^2)p_1 + f_2'(\omega^2)p_2 + f_0'(\omega^2)] = 0 \end{cases} \quad \omega > 0 \quad (29)$$

where $f_i'(\cdot)$, $i = 0, 1, 2$ denote first derivatives with respect to ω . Polynomial functions f_0, f_1, f_2 have an easy interpretation in terms of the frequency response $G(j\omega)$. In particular, setting

$$G(j\omega) = \frac{N_r(j\omega) + jN_i(j\omega)}{D_r(j\omega) + jD_i(j\omega)} \quad (30)$$

it is easy to show that

$$\begin{aligned} f_0(\omega^2) &= N_r(j\omega)D_r(j\omega) + N_i(j\omega)D_i(j\omega) \\ f_1(\omega^2) &= \omega[N_r(j\omega)D_r(j\omega) - N_i(j\omega)D_i(j\omega)] \\ f_2(\omega^2) &= [D(j\omega)]^2 \end{aligned} \quad (31)$$

The above equations allow one to give a graphical interpretation of the critical points in terms of the Popov locus in the complex plane (Siljak, 1969). In the parameter plane each curve of the envelope is a straight line. As a consequence, the set $S_1(p_2)$ may be readily obtained by computing the real positive solutions ω_i^2 of the polynomial equation

$$f_1(\omega_i^2) = 0 \quad (32)$$

and then computing the corresponding solutions for p_2

$$p_2 = -f_0(\omega_i^2)/f_2(\omega_i^2) \quad (33)$$

Notice that from the second equation in (31) solution frequencies ω_i correspond to points where the Popov locus crosses the real axis.

Computation of the solutions belonging to the critical set $S_2(p_2)$ does not simplify significantly in the special case $p_2 = k^{-1}$. These solutions correspond to intersections of the real axis with straight lines admitting possible multiple tangency points with the Popov plot. In fact, these solutions represent critical points of the curve $R_{\Omega}(p_1, p_2) = 0$, i.e. points such that there exist at least two different values of frequency ω satisfying the envelope equations (29).

The set $S_1(p_2)$ can be readily computed because the last equations of (22, 23) are linear in p_1, p_2 , so that the solution of each of the two systems requires to solve one polynomial equation. From a graphical point of view, these points correspond to lines passing through one of the end points of the Popov plot and tangent to it at that point and/or some other point(s).

Observation 3. The feedback gain interval $[0, k_H]$ for which the linear system $G(s)$ is closed loop asymptotically stable is usually called *Hurwitz sector* for system (1). It can be easily checked that if $p_2'' \in S_1(p_2)$, then $p_2'' = k_H^{-1}$, i.e. the Popov sector coincides with the Hurwitz sector. Hence, the Popov criterion allows one to conclude that the well known Aizerman conjecture is true.

An example of computation of the maximum k for which the Lur'e system (1) is AS according to the Popov criterion for a classical example taken from the literature is given in Section 5.

4. Robust absolute stability for planar uncertainty sets

In this section we assume that the sector bound k in (2) is given. We consider the case in which the components of vector $p \in \mathbb{R}^2$ are uncertain physical parameters affecting polynomially the coefficients of $G(s; p)$ (or the entries of (A, b, c)). Let the uncertainty set be defined as a rectangle in the parameter plane

$$U_p = \{p \in \mathbb{R}^2 : p_i' \leq p_i \leq p_i'', i = 1, 2\}. \quad (34)$$

Assuming that $\forall p \in U_p$ the linear part of (1) is asymptotically stable and that U_p contains at least one point $p = p^0$ for which $G(s; p)$ satisfies the Popov criterion, absolute stability can be studied in two successive steps. In a first step we ascertain if the boundary ∂U_p intersects the envelope $E(p)$. In the second step it is checked if there exist points of the envelope in the interior of U_p .

Step 1. We look for possible intersections of the rectangle sides with the envelope $E(p)$ given in (13–16). Considering

sides of the rectangle, i.e. setting alternatively $p_i = p_i'$ (or $p_i = p_i''$), $i = 1, 2$, possible intersections can be computed by applying the procedure given in Section 3 for each side.

Step 2. To check for absence of envelope points inside U_p , we have to verify that the polynomial equation systems (22) and (23) have no solutions belonging to U_p . Moreover, it must be checked that the following polynomial equation system

$$\begin{cases} R_{\Omega}(p_1, p_2) = 0 \\ \partial R_{\Omega}(p_1, p_2) / \partial p_1 = 0 \\ (\text{or } \partial R_{\Omega}(p_1, p_2) / \partial p_2 = 0) \end{cases} \quad (35)$$

has no real solution in U_p . Notice that if we denote by $S_u(p)$ the set of solutions of (35), only the following subset must be considered

$$S_u^L(p) = \{S_u(p) \cap E(p)\}. \quad (36)$$

From a practical point of view, system (35) can be solved, as shown for other cases in Section 3, by applying the resultant theorem. The set $S_u^L(p)$ can be computed immediately selecting solutions $p \in S_u(p)$ for which there exists some positive real value ω solving simultaneously both equations (9).

As a last observation on Step 2, it is worth noting that if for a fixed value of the Popov parameter the test fails, i.e. the set $S_u^L(p)$ is not empty, we have to perform the test for different values of θ . Hence, in general it may happen that Step 2 must be repeated for all real values θ , meaning that the solution of a family of problems like that solved above may be needed.

As a final comment, we notice that since the Popov criterion is only sufficient for AS, the fact that the test proposed fails in assessing AS of an uncertainty set U_p in general does not allow to conclude that system (1) is *not* AS for $p \in U_p$. A negative answer of the test would mean only that the Popov criterion does not allow one to assess robust absolute stability, so that other alternative criteria should be employed.

5. Numerical examples

In this section we present two examples showing applications of the results presented in previous sections. In the first example, a well known system is considered and the maximum Popov sector is computed. In the second example, we determine the maximum allowable uncertainty domain for a scalar parameter affecting the linear part of a Lur'e system, for a prescribed class of nonlinear functions \mathcal{F}_k .

Example 1 (Narendra and Taylor, 1973, p. 173; Safonov and Weytznar, 1987). Consider a Lur'e system where the linear part is described after a pole-shifting by

$$G(s) = \frac{3(s+1)}{s^4 + s^3 + 25s^2 + 3s + 3}. \quad (37)$$

From standard arguments it follows that the Munoitz sector is $[0, 7)$. Thus, by defining $(p_1, p_2) = (\theta, k^{-1})$, we observe that we need only consider values p_2 in the interval $(0, 14285, \infty]$. Moreover, it can be easily checked from the structure of $P(\omega^2; p)$ that the point $(p_1, p_2) = (0, \infty)$ satisfies the Popov condition (4), so that we can assume as nominal value $p_2^0 = \infty$. The functions $f_i(\cdot)$, $i = 0, 1, 2$ in (31) are given by

$$\begin{cases} f_0(\omega^2) = -66\omega^2 + 9 \\ f_1(\omega^2) = -3\omega^6 + 72\omega^4 \\ f_2(\omega^2) = \omega^8 - 49\omega^6 + 625\omega^4 - 141\omega^2 + 9. \end{cases} \quad (38)$$

From equations (32) and (33) we obtain the set

$$S_1(p_2) = \{-1, 0.14285\}. \quad (39)$$

The envelope equations are

$$\begin{aligned} F_1(\Omega; p) &= p_2\Omega^4 - (49p_2 + 3p_1)\Omega^3 + (625p_2 + 72p_1)\Omega^2 \\ &\quad - (141p_1 + 66)\Omega + 9p_1 + 9 = 0 \\ F_2(\Omega; p) &= 4p_2\Omega^3 - 3(49p_2 + 3p_1)\Omega^2 + 2(625p_2 + 72p_1)\Omega \\ &\quad + (141p_1 + 66) = 0. \end{aligned} \quad (40)$$

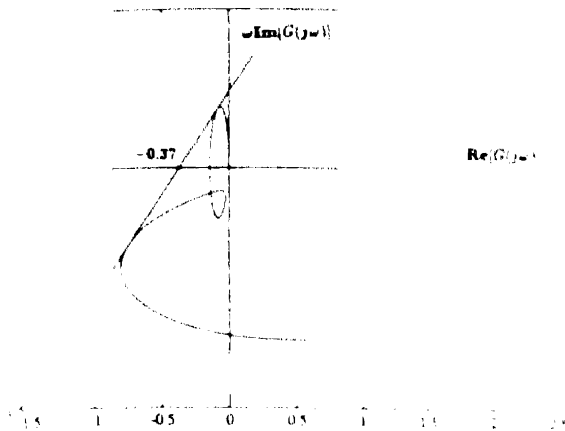


FIG. 1. Popov line for maximal absolute stability sector of Example 1.

Since the degree of polynomial $F_1(\Omega; p)$ in Ω is 4, from Observation 2 of Section 3, we can expect that self intersection points of the envelope may exist. In fact, computing the resultant $R_\Omega(p_1, p_2)$ and solving (21), we obtain the set

$$S_2(p_2) = \{-1.443, 0.37\} \quad (41)$$

The set $S_1(p_2)$ obtained by solving equations (22) and (23) is given by

$$S_1(p_2) = \{-1, 0\} \quad (42)$$

From (39), (41) and (42) we obtain

$$S(p_2) = \{-1.443, -1, 0, 0.14285, 0.37\} \quad (43)$$

The extremal set $S_*(p_2)$ is obtained by checking properties Π_1 and Π_2

$$S_*(p_2) = \{0.37\}$$

Hence, we select $p_2^m = 0.37$ and $p_2^M = \infty$ obtaining that the considered system is AS for $k \in [0, 2.7)$. The corresponding values of θ and ω for $k = 2.7$ are found to be $\theta = 0.781$ and $\omega_1 = 1.54$, $\omega_2 = 5.24$. These two values of ω correspond to the tangency points of the straight line defined by $k = 2.7$ and $\theta = 0.781$ with the Popov locus (see Fig. 1).

Example 2. Consider the system described by the transfer function

$$G(s; p_2) = \frac{11}{(s+1)(s^2 + p_2 s + 9)} \quad (44)$$

where p_2 is an uncertain parameter with nominal value $p_2^0 = 10$ and the nonlinearity $f(\cdot) \in \mathcal{F}_k$ with $k = 1$. From standard arguments, we obtain that the closed loop system with a linear constant feedback with gain $k = 1$ is asymptotically stable for $p_2 \in (1, \infty)$. As in Section 3, we set $p_1 = \theta$. The functions $f_i(\cdot)$, $i = 0, 1$ in (17) are given by

$$\begin{cases} f_0(\omega^2, p_2) = \omega^6 + (p_2^2 - 17)\omega^4 + (52 - 11p_2 + p_2^2)\omega^2 + 180 \\ f_1(\omega^2, p_2) = -11\omega^4 + 11(9 + p_2)\omega^2 \end{cases} \quad (45)$$

From equations (19) we obtain the set

$$S_1(p_2) = \{0, 1\} \quad (46)$$

The envelope equations are

$$\begin{cases} F_1(\Omega; p) = \Omega^3 + (p_2^2 - 11p_1 - 17)\Omega^2 \\ \quad + (p_2^2 + 11p_1 p_2 - 11p_2 + 99p_1 + 52)\Omega + 180 = 0 \\ F_2(\Omega; p) = 3\Omega^2 + 2(p_2^2 - 11p_1 - 17)\Omega \\ \quad + (p_2^2 + 11p_1 p_2 - 11p_2 + 99p_1 + 52) = 0. \end{cases} \quad (47)$$

From Observation 2 in Section 3, we notice that since the degree of the polynomial $F_1(\Omega; p)$ in Ω is 3, the only possible points belonging to the set $S_2(p_2)$ must necessarily yield a triple positive solution for Ω in $F_1(\Omega; p)$. This cannot happen because $a_n(p) = 180 > 0$. Therefore, the set $S_2(p_2)$ is empty. This can be also verified by computing the resultant $R_\Omega(p_1, p_2)$ and solving (21).

Also the set $S_3(p_2)$ turns out to be empty because the coefficients $a_1(p)$ and $a_2(p)$ never vanish. Consequently, the set $S(p_2)$ is equal to $S_1(p_2)$ and $S_*(p_2) = \{1\}$. Hence, we obtain $p_2^m = 1$ and $p_2^M = \infty$, obtaining that the considered Lur'e system is AS for $p_2 \in (1, \infty)$. The corresponding values of θ and ω for $p_2 = 1$ are $\theta = 0.2$ and $\omega = 3.162$.

6. Conclusions

In this paper an analytical method is proposed for robust absolute stability analysis of Lur'e control systems subject to parameter variations. The case where system perturbations are due to one or two uncertain parameters has been considered and a method is given which allows one to estimate maximal domains of absolute stability in parameter space, based on the Popov criterion. Further research is needed in the direction of studying robust absolute stability by means of different criteria, which may take into account more information about the feedback nonlinear function than the sector condition. Moreover, effective methods of analysis for problems involving several physical parameters appear to be of primary interest for future work.

Acknowledgements—This work was partially supported by funds of Ministero della Università e della Ricerca Scientifica e Tecnologica.

References

- Bhattacharyya, S. P. (1987), *Robust Stabilization Against Structured Perturbations*, Lecture Notes in Control and Information Sciences 99, Springer, Berlin.
- Bose, N. K. and J. E. Delansky (1989), Boundary implications for interval positive rational functions, *IEEE Trans. Circ. Syst.* **CAS-36**, 454–458.
- Genesio, R. and A. Tesi (1988), Results on the stability robustness of systems with state space perturbations, *Syst. Control Lett.* **11**, 39–47.
- Jacobson, N. (1964), *Lectures in Abstract Algebra*, Vol. III, Von Nostrand, Princeton, NJ.
- Narendra, K. S. and J. H. Taylor (1973), *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York.
- Neimark, Yu. I. (1949), *Stability of Linearized Systems* (in Russian), LKVVIA, Leningrad.
- Popov, V. M. (1962), Absolute stability of nonlinear systems of automatic control, *Aut. Remote Control*, **22**, 857–875. (Russian original published in 1961).
- Safonov, M. G. and G. Wertzner (1987), Computer-aided stability analysis renders Popov criterion obsolete, *IEEE Trans. Aut. Control*, **AC-32**, 1128–1131.
- Sideris, A. and R. S. S. Peña (1988), Fast computation of the multivariable stability margin for real interrelated uncertain parameters, *Proc. ACC*, Atlanta, GA.
- Siljak, D. D. (1969), *Nonlinear Systems: The Parameter Analysis and Design*, Wiley, New York.
- Siljak, D. D. (1971), New algebraic criteria for positive realness, *J. Franklin Inst.*, **291**, 109–120.
- Siljak, D. D. (1989a), Parameter space methods for robust control design: A guided tour, *IEEE Trans. Aut. Control*, **AC-34**, 674–688.
- Siljak, D. D. (1989b), Polytopes of nonnegative polynomials, *Proc. ACC*, Pittsburgh, PA.
- Tesi, A. and A. Vicino (1988), Robustness analysis for uncertain dynamical systems with structured perturbations, *Proc. 27th Conf. on Decision and Control*, Austin, TX, 519–525. Also *IEEE Trans. Aut. Control* 1990, **AC-35**, 191–195.
- Vicino, A., A. Tesi and M. Milanese (1988), An algorithm for nonconservative stability bounds computation for systems with nonlinearly correlated parametric uncertainties, *Proc. 27th IEEE Conf. on Decision and Control*, Austin, TX, 1761–1766. Also *IEEE Trans. Aut. Control* 1990, **AC-35**, 835–841.
- Vicino, A. (1989), Maximal polytopic stability domains in parameter space for uncertain systems, *Int. J. Control*, **49**, 351–361.
- Voronov, A. A. (1989), On improving absolute stability criteria for systems with monotonic nonlinearities and the method of absolute stability regions construction, *Preprints IFAC Symp. Nonlinear Control Systems Design*, Capri, Italy, pp. 231–235.

Brief Paper

On the Robustness of Discrete-time Indirect Adaptive (Linear) Controllers*

F. GIL^{†‡}, M. M'SAAD[†], J. M. DION[†] and L. DUGARD[‡]

Key Words— Adaptive control; convergence analysis; stability; robustness.

Abstract—Remarkable research activity has been devoted to the design of certainty equivalence-based adaptive controllers which perform well in those ubiquitous non-idealities as bounded disturbances, time varying parameters and some classes of unmodeled dynamics. The long standing issue is the estimated model admissibility condition, i.e. the underlying linear control law should stabilize the estimated model. The motivation of this paper is to propose a general framework for robustly designing the adaptive linear controllers, irrespective of the underlying control law. More specifically a new solution to the problem of the estimated model admissibility using an *ad hoc* modification of the control law is given. Such a modification consists in adding an internal impulse exciting sequence, while freezing the controller parameters, whenever the estimated model admissibility is lost.

1. Introduction

INTENSIVE RESEARCH developments over the last decades have shown that the flexibility provided by the adaptive control is remarkably broad. The seminal results within this context were specifically concerned with direct adaptive controllers, i.e. model reference as well as single stage optimal stochastic adaptive controllers. Complete stability and convergence results were obtained in this context under ideal conditions: the plant to be controlled is completely described by a linear time invariant system of known structure, and the allowable disturbances are those moving average of independent zero mean random variables [see Fuchs (1982) and Goodwin and Sin (1984) for an elegant presentation]. Although these results provided better theoretical understanding of the adaptive control concept and led to certain successful industrial applications (Åström, 1987), their practical implications have been shown to be questionable. Indeed it has been shown that the resulting closed loop adaptive systems could be made unstable in the presence of bounded disturbances (Egardt, 1979), time varying parameters (Anderson and Johnstone, 1983) and reduced order modeling (Ioannou and Kokotovic, 1983; Rohrs *et al.*, 1985). Moreover, the minimum phase design assumption involved in direct adaptive controllers is more an exception than a rule in the discrete-time context (Åström *et al.*, 1984).

The major subsequent developments in adaptive control theory have been devoted to the problem of preserving the stability of the adaptive controllers in spite of bounded external disturbances as well as time varying and unmodeled dynamics. Two main approaches have been adopted to deal with such a problem. The first one relies upon the importance of the *concept of persistent excitation* in adaptive identification and control [see Narendra and Annaswamy (1987) and references list therein]. More specifically, the adaptive controllers derived in ideal situations have been shown to be locally stable in realistic situations provided that some of the internal signals are dominantly rich. These stability results, local in nature, provided new insights about the local behaviour of adaptive controllers in non-ideal situations, namely sharp bounds for local stability instability domains have been given (Anderson *et al.*, 1986). It is however worth mentioning that fundamental design questions, such as how to ensure the involved persistent excitation, and how to enlarge the local stability domain, have not yet been solved. The second approach consists of redesigning the parameter adaptation algorithms to accommodate bounded disturbances, time varying and unmodeled dynamics. This leads to what is referred to as *robust parameter estimators* which, besides improving the stability robustness, allow the incorporation of certain prior information regarding the plant to be controlled into the design, e.g. sharp bound on external disturbances, allowable convex domain in the parameter space. This ends the era of the 'black box' approach, allowing any good engineering intuition as well as physical insight to be used.

Three modifications are commonly used in the robust adaptive control literature. The dead zone first proposed by Egardt (1979) as well as Peterson and Narendra (1982) consists of conditional updating based on prior information regarding sharp bound on the disturbances sequence. Although such a modification provides greater flexibility in stability analysis (Samson, 1983a; Middleton *et al.*, 1988; Middleton and Goodwin, 1988; Giri *et al.*, 1987, 1988a), its practical significance is indeed questionable, particularly in the presence of unmodeled and time varying dynamics. Indeed the choice of the relative dead zone size is by no means obvious. The second modification suggested by Egardt (1979) as well as Kreisselmeier and Narendra (1982) consists of projecting the parameter estimates in a certain allowable region of the parameter space by using prior knowledge about the plant to be controlled. The latter prior information is generally easier to get than that involved in dead zone modification. The third modification is due to Ioannou and Kokotovic (1983). It consists of contracting the parameter estimates and is generally referred to as a *modification or leakage*. All these modifications have been introduced to prevent the parameter estimates drift caused by bounded disturbances due to the integral nature of the original adaptive laws. They are however not sufficient to deal with the ubiquitous problem of unmodeled dynamics as their corresponding disturbances cannot be *a priori* assumed bounded any longer. Such a problem has been addressed by Praly (1982, 1983) who introduces an appropriate data

* Received 10 August 1987; revised 15 June 1989; revised 3 January 1990; received in final form 5 March 1990. The original version of this paper was presented at the IFAC Workshop on Robust Adaptive Control, Newcastle, Australia, August 1988. This paper was recommended for publication in revised form by Associate Editors G. Kreisselmeier and M. Y. Mareels under the direction of Editor P. C. Parks.

[†]Laboratoire d'Automatique, Grenoble (J. A. C.N.R.S., 228), E.N.S.T.G., E.N.P.G., B.P. 46, 38402 Saint Martin d'Hères, France.

[‡]Now with the Laboratoire d'Automatique et d'Informatique Industrielle (I.A2I), Ecole Mohammadia d'Ingénieurs, Rabat, Morocco.

normalization that turns the problem of unbounded disturbances into a bounded disturbances one. This makes it possible to use the above cited modifications in the presence of unmodeled dynamics.

On the other hand, a useful approach for studying the stability of indirect adaptive controllers has been proposed by De Larminat (1981), Fuchs (1980) and Samson (1982) for the ideal case and extended to more realistic situations by De Larminat (1986), Egardt and Samson (1982) and Samson (1983a, b). This legitimates the adaptive control of non-minimum phase plants with unknown and possibly varying time-delay. The corner stone of such an approach is that any linear control law can be combined with a parameter estimator which satisfies a well defined set of properties to provide a globally stable indirect adaptive linear controller. The only stability property which is not necessarily provided by the available parameter estimators is what is referred to, in De Larminat (1984), as the *estimated model admissibility*: the underlying linear control law should stabilize the estimated plant model.

Three methods have been proposed in the literature for achieving the estimated model admissibility condition. The first one consists of simply restricting the parameter space to a convex admissible domain (Goodwin and Sin, 1984; Kreisselmeier, 1985, 1986a; Middleton *et al.*, 1988). The underlying philosophy of the second approach is as follows: since the plant under control is assumed to be admissible, incorporate an *ad-hoc* procedure to avoid non-admissible estimated models (De Larminat, 1984; Lozano and Goodwin, 1985). The third approach aims to ensure the convergence of the parameter estimates (close) to their true values. Such an objective has been particularly achieved for pole placement adaptive controllers in an ideal case, using an appropriate exciting sequence (Elliot *et al.*, 1985; Goodwin and Teoh, 1985; Goodwin *et al.* 1985; Anderson and Johnstone, 1985; Kreisselmeier and Smith, 1986; Polderman, 1989).

A new approach to provide an *ad-hoc* richness property in a non-ideal situation has been proposed in Giiri *et al.* (1987, 1988a). To this end, the adaptive control law is modified such that an appropriate internal impulse exciting sequence is added while the controller parameters are frozen over a suitable time horizon, whenever the plant model admissibility is lost. The underlying results are however limited from two points of view. Firstly the plant model is assumed to be time invariant and the allowable unmodeled dynamics are those which are always small with respect to the input-output signals. Secondly the stability analysis is specifically derived for pole placement adaptive controllers with dead-zone. The main motivation of this paper is to generalize the above investigations in two directions. (1) The class of the plants to be controlled is enlarged to include both time-varying parameters and unmodeled dynamics. The effects of these non-idealities are assumed to be asymptotically small only in the mean. (2) The stability analysis is performed irrespective of both the control and adaptation laws.

The paper is organized as follows. In Section 2, the problem formulation is stated to emphasize the design assumptions. The considered class of adaptive controllers is presented in Section 3. Section 4 is devoted to the closed loop stability analysis. First the proposed adaptive control law is shown to guarantee the required plant model admissibility. Then the uniform boundedness of the closed loop signals is established using a general stability lemma. Some concluding remarks end the paper.

2. Problem statement

The class of the plants to be controlled is first defined. Then a unified linear control law structure is given for versatility purpose. Finally, the control objective is stated.

Throughout the paper, a real sequence $\{x(t)\}$ will be said *μ -asymptotically small in the mean (μ -ASM)* if

$$\limsup \limsup \sum_{t=0}^k s(t).$$

The set of all these sequences will be noted $S_\mu(\mu)$. The definition of (μ -ASM) sequences allows compact statement of μ -small in the mean sequences (Praly, 1982).

2.1. *The assumed plant model.* We shall consider the class of plants whose input-output behaviour is "almost" like a low order linear time varying model, which can be described in the discrete time context as follows.

$$\begin{aligned} A(\theta^*(t), q^{-1})z(t) &= u(t) \\ y(t) &= B(\theta^*(t), q^{-1})z(t) + \eta(t) \end{aligned} \quad (2.1)$$

with

$$\begin{aligned} \theta^*(t) &= [a_1^*(t) \cdots a_n^*(t)b_1^*(t) \cdots b_n^*(t)]^T \\ A(\theta^*(t), q^{-1}) &= 1 + a_1^*(t)q^{-1} + \cdots + a_n^*(t)q^{-n} \\ B(\theta^*(t), q^{-1}) &= b_1^*(t)q^{-1} + \cdots + b_n^*(t)q^{-n} \end{aligned} \quad (2.2)$$

where $u(t)$, $y(t)$ and $z(t)$ denote the plant input, output and partial state, respectively; $\{\eta(t)\}$ is a disturbance sequence which incorporates all factors affecting the plant output; q^{-1} is the backward shift operator (i.e. $q^{-1}x(t) = x(t-1)$).

We assume that the integer n is chosen *a priori* so that the corresponding sequences $\{\theta^*(t)\}$ and $\{\eta(t)\}$ satisfy:

A1. There exists a known scalar R_μ such that

$$\|\theta^*(t)\| < R_\mu$$

A2. There exists a positive scalar δ such that

$$\{-|\det M_r(A(\theta^*(t), q^{-1}), B(\theta^*(t), q^{-1}))|\} \in S_\mu(-\delta)$$

where $M_r(X(q^{-1}), Y(q^{-1}))$ denotes the Sylvester resultant associated with the $X(q^{-1})$ and $Y(q^{-1})$ polynomials.

A3. There exists a positive scalar ν such that

$$\{|\eta(t)|/m(t)\} \in S_\mu(\nu)$$

with

$$m(t) = \sigma m(t-1) + \max\{|\phi(t)|, m_0\}$$

$$0 \leq \sigma < 1, \quad m_0 > 0, \quad m(0) > 0$$

and

$$\phi(t) = [-y(t-1) \cdots -y(t-n), u(t-1) \cdots u(t-n)]^T.$$

It is worth mentioning that the plant model parameters are allowed to be time varying. Such a feature is more an exception than a rule in the available literature, though the problem of time varying dynamics is the prime motivation of the adaptive control concept. The assumption A2 means that the plant model is sufficiently controllable in the mean; this is coherent with the time varying feature. The assumption A3 means that the unmodeled response $\{\eta(t)\}$ should be linearly bounded, in the mean, by the input-output data activity measure $\{m(t)\}$. This characterizes a relatively important class of plant-model mismatch, namely the neglected high order terms and those non-linear characteristics that can be linearly dominated by the input-output signals (Praly, 1982). Notice that the sequence $\{|\eta(t)|/m(t)\}$ is uniformly bounded.

2.2. *The controller structure.* As our aim is to provide a general approach for achieving the stability of adaptive linear controllers irrespective of the underlying linear control design, we consider the following control law structure. (In the remainder of the paper the argument q^{-1} will be omitted).

$$R(\theta^*(t))u(t) + S(\theta^*(t))y(t) = T(\theta^*(t))y^*(t) \quad (2.3)$$

with

$$\begin{aligned} R(\theta^*(t), q^{-1}) &= 1 + r_1^*(t)q^{-1} + \cdots + r_n^*(t)q^{-n-1} \\ S(\theta^*(t), q^{-1}) &= s_n^*(t) + s_{n-1}^*(t)q^{-1} + \cdots + s_1^*(t)q^{-n-1} \\ T(\theta^*(t), q^{-1}) &= t_0^*(t) + t_1^*(t)q^{-1} + \cdots + t_n^*(t)q^{-n} \end{aligned} \quad (2.4)$$

where $\{y^*(t)\}$ is a user-specified uniformly bounded reference sequence. The $R(\theta^*(t), q^{-1})$, $S(\theta^*(t), q^{-1})$ and $T(\theta^*(t), q^{-1})$ are polynomials which are evaluated according to the considered control objective, e.g. model reference for invertibly stable systems, pole placement, and receding

horizon linear quadratic control (See Kreisselmeier, 1985, 1986a, b; M'Saad *et al.*, 1985 and Bitmead *et al.*, 1988) for more details).

In the following, we will introduce the concept of plant mode admissibility with respect to the control law (2.3)–(2.4). To this end, let $K(\cdot)$ be the underlying evaluation function of the control law (2.3), i.e.

$$K: \mathbb{R}^{2n} \rightarrow \mathbb{R}^m$$

$$\theta \rightarrow K(\theta) = [r_1(\theta) \cdots r_{n-1}(\theta) \lambda_0(\theta) \cdots \lambda_{n-1}(\theta) \lambda_0(\theta) \cdots \lambda_n(\theta)]^T$$

Definition 2.1. $\mathcal{X}_a \subset \mathbb{R}^{2n}$ is said to be an admissible domain with respect to $K(\cdot)$ if there exists $\rho \in [0, 1]$ such that

- (i) For any $\theta \in \mathcal{X}_a$, if $A(\theta, q^{-1})R(\theta, q^{-1}) + B(\theta, q^{-1})S(\theta, q^{-1}) = 0$ then $\|q\| \leq \rho$
- (ii) $K(\cdot)$ is Lipschitz on \mathcal{X}_a .

That is, the admissible domain with respect to $K(\cdot)$ contains the admissible plant models, i.e. the parameter vectors θ , that could be sufficiently stabilized by the control law $K(\cdot)$. In order to get a well-posed control problem, the plant model (2.1)–(2.2) together with the control law (2.3)–(2.4) should satisfy the following assumptions.

A4. $(\forall t, \theta^*(t) \in \mathcal{X}_a)$ and $(\exists \epsilon > 0)(\forall t) d(\theta^*(t), \partial \mathcal{X}_a) \geq \epsilon$ where $d(\cdot, \cdot)$ is a given distance and $\partial \mathcal{X}_a$ is the \mathcal{X}_a boundary.

It is worth noticing that \mathcal{X}_a is not required to be explicitly known. A condition for belonging to \mathcal{X}_a will be sufficient e.g. the admissibility condition is reduced to a coprimeness test of the polynomial sequence pair $(A(\theta^*(t)), B(\theta^*(t)))$ when a pole placement control is considered. The assumption A4 stipulates that the plant model under test is (uniformly) admissible with respect to the control law.

2.3 The control objective Let v be the least non-negative scalar so that

$$\{\|\theta^*(t) - \theta^*(t-1)\|\} \in S_0(v) \quad (2.5)$$

v exists according to the assumption A1. And letting

$$\mu = \max(v, \nu) \quad (2.6)$$

where ν is the non-negative scalar defined in assumption A3. The control objective consists in designing an indirect adaptive controller, irrespective of the control law as discussed above, for the plant (2.1)–(2.2) subject to assumptions A1–A4 so that

- (i) there exists a non-negative scalar μ^* such that if $\mu \in [0, \mu^*]$ then all the closed loop signals are uniformly bounded;
- (ii) the performances of the underlying control law are asymptotically achieved in the ideal case (i.e. $\theta^*(t) = \theta^*(t-1)$ and $\eta(t) = 0$ for all t).

Such a result will be of prime importance from both stability robustness and performance points of view. Indeed the involved uniform boundedness will be obtained in spite of a relatively large class of unmodeled dynamics, as pointed out in assumption A3, and in spite of time varying plant model parameters. The latter incorporates jump as well as drift parameters. "Drift parameters" refers to the case where the parameters frequently undergo small variations whereas the "jump parameters" means that the parameter could change largely but infrequently. Such a class of time-varying systems has been recently investigated by Kreisselmeier (1986b), Middleton and Goodwin (1988), De Larminat (1986) and Tsakalis and Ioannou (1987).

3. The adaptive control algorithm

In this section we will first specify the properties that should be satisfied by the parameter adaptation algorithm, bearing in mind the considered class of plants. Then the adaptive control law is given with a particular emphasis on the new features

3.1 Robust parameter adaptation algorithm The robust parameter adaptation algorithm is used to generate a parameter estimates sequence $\{\hat{\theta}(t)\}$ of the unknown parameters sequence $\{\theta^*(t)\}$. The word "robust" means that the parameter adaptation algorithm is able to accommodate bounded disturbances, time varying parameters and unmodeled dynamics. In the following, we will describe the robust stability properties which determine the choice of the parameter adaptation algorithm for its successful use in adaptive control.

P1. $\{\hat{\theta}(t)\}$ is uniformly bounded

P2. There exists a positive constant K_μ (independent of μ) such that

$$\|\hat{y}(t)\| \in S_0(K_\mu \mu^*) \text{ with } \hat{y}(t) = y(t) - \hat{\theta}(t)^T \phi(t)$$

where $\|\cdot\|$ denotes data normalization by the sequence $\{m(t)\}$ defined in assumption A3 and a is a positive constant, independent of μ .

P3. There exists a positive constant K_μ (independent of μ) such that

$$\|\hat{\theta}(t) - \hat{\theta}(t-1)\| \in S_0(K_\mu \mu^*)$$

These properties are analogous to those given in De Larminat (1981), Fuchs (1980) and Samson (1982) for the ideal case as well as in Egardt and Samson (1982) and Samson (1983a) for the case of bounded disturbances. Similar stability properties have been required by Praly (1982), Samson (1983b), De Larminat (1986) and De Larminat and Raynaud (1988) to deal with the same control problem. The uniform boundedness property P1 excludes the possibility of parameter estimates wind up. Although this property is easily verified in the ideal case, it may be lost in non-ideal situations, particularly when the signals involved in the adaptive loop are not persistently exciting (Anderson *et al.*, 1986). More specifically, such a property was the prime motivation of the adaptive law modifications described earlier in the introduction. On the other hand the properties P2 and P3 ensure that the normalized adaption error and the parameter estimates variations are sufficiently small in the mean. The involved "smallness" is related to the tolerable unmodeled and time varying dynamics, i.e. the scalar μ . Notice that, all these properties are satisfied by the available robust parameter estimation algorithms (Egardt, 1979; Praly, 1982; De Larminat and Raynaud, 1988), e.g. regularized normalized least squares with parameter projection or contraction.

3.2 The indirect adaptive linear control law A robust indirect adaptive linear controller is simply obtained by combining a parameter estimator satisfying the robust properties P1–P3 with the certainly equivalence form of the linear control law (2.3)–(2.4), i.e.

$$R(\hat{\theta}(t))u(t) + S(\hat{\theta}(t))y(t) = T(\hat{\theta}(t))y^*(t) \quad (3.1)$$

The resulting adaptive controller is robustly stable provided that the estimated plant model is sufficiently admissible with respect to the underlying control law. More precisely, the parameter estimates sequence should satisfy the following property.

P4. There exists a subsequence $\{t_k\}_{k \in \mathbb{N}}$ such that

$$\{\hat{\theta}_k - \hat{\theta}_{k-1}\} \text{ is uniformly bounded and } (\forall k \in \mathbb{N}) \theta(t_k) \in \mathcal{X}_a$$

Notice that $\hat{\theta}(t)$ is not required to belong to \mathcal{X}_a for all t .

The main problem is that property P4 is not necessarily satisfied by the available parameter estimators. Several solutions have been proposed to deal with such a problem as pointed out in the introduction. In this paper, the *estimated model admissibility* is ensured by modifying the adaptive control law (3.1) as follows

$$R(\hat{\theta}_n(t))u(t) + S(\hat{\theta}_n(t))y(t) = T(\hat{\theta}_n(t))y^*(t) + \alpha(t) + \beta(t) \quad (3.2)$$

where the sequences $\{\hat{\theta}_n(t)\}$, $\{\alpha(t)\}$ and $\{\beta(t)\}$ are evaluated as follows.

Letting l the integer subset given by

$$l = \{t \in \mathbb{N} : \hat{\theta}(t) \notin \mathcal{X}_a \text{ and } |t - t_{k-1}| \geq 4n + 1\}$$

If there exists $t \in N$ such that $t \in [t_i, t_i + 4n - 2]$ then

$$\hat{\theta}_a(t) = \hat{\theta}(t_i) \text{ with } t_i = \max \{k : k < t_i \text{ and } \hat{\theta}(k) \in \mathcal{A}_a\} \quad (3.3a)$$

$$\alpha(t) = 2n \left(\sum_{j=0}^{n-1} |t_j(\hat{\theta}_a(t))| \right) v^* \quad \left. \vphantom{\sum} \right\} \text{ for } t = t_i + 2n - 1 \quad (3.3b)$$

$$\beta(t) = \beta m(t) \quad (3.3c)$$

$$\alpha(t) = 0 \quad (3.3d)$$

$$\beta(t) = 0 \quad \text{for } t \neq t_i + 2n - 1 \quad (3.3e)$$

Else

$$\hat{\theta}_a(t) = \hat{\theta}(t) \quad (3.3f)$$

$$\alpha(t) = 0 \quad (3.3g)$$

$$\beta(t) = 0 \quad (3.3h)$$

where $\{\hat{\theta}(t)\}$ satisfies properties P1-P3, $t_i(\hat{\theta}_a(t))$ are the coefficients of the polynomial $T(\hat{\theta}_a(t))$, v^* is a finite upper bound on the reference sequence $\{y^*(t)\}$, and β is an arbitrary chosen positive scalar. Note that $\hat{\theta}_a(0)$ has to be admissible, i.e. $\hat{\theta}_a(0) \in \mathcal{A}_a$.

In order to give some insights about the adaptive control law modification, let t_i be the first instant when $\hat{\theta}(t)$ does not belong to \mathcal{A}_a . ($\hat{\theta}(t_i)$ is not admissible). Then, $\hat{\theta}_a(t)$ is chosen as $\hat{\theta}(t_i)$, $t_i = t_i - 1$, for $t \in [t_i, t_i + 4n - 2]$ and an internal impulse signal $[\alpha(t) + \beta(t)]$ given by 3.3b and 3.3c is added at time $t_i + 2n - 1$. Let us consider the first instant $t_i < t_{i+1} + 4n - 1$ such that $\hat{\theta}(t_i)$ does not belong to \mathcal{A}_a . Then, $\hat{\theta}_a(t)$ is chosen for any $t \in [t_i, t_i + 4n - 2]$ as the last admissible estimated parameter vector $\hat{\theta}(t_i)$, where t_i is defined in (3.3a), and an internal impulse is added at time $t_i + 2n - 1$.

Notice that $\{\hat{\theta}_a(t)\}$ is constructively admissible, i.e. $\hat{\theta}_a(t) \in \mathcal{A}_a$ for all t , even if $\hat{\theta}(t)$ is not always admissible.

Moreover the controller parameters are frozen over the time interval $[t_i, t_i + 4n - 2]$. The internal impulse signal provides then the necessary amount of excitation to ensure the estimated plant model admissibility as we will show.

Although the controller parameter freezing as well as the exciting internal signal $(\alpha(t) + \beta(t))$ may be viewed as technical devices to deal with the admissibility problem, they may have a quite interesting interpretation. Indeed the excitation of the signals involved in an adaptive control system is due to three different sources, namely the reference sequence, the controller parameter variations and the unmodeled response. However it is well known that these natural exciting effects may cancel each other, leading to a non-admissible accumulation point in the parameter space. The underlying philosophy of the proposed approach consists in compensating these effects and then producing an appropriate excitation amount which ensures the estimated plant model admissibility according to the property P4. More precisely the reference signal is compensated by the sequence $\{\alpha(t)\}$, the controller parameter variations are completely removed by the freezing process, the unmodeled response effects are cancelled by the sequence $\beta(t)$. The latter produces also the required amount of excitation. The use of internal exciting signals to cope with the estimated model admissibility problem has been first suggested in Kreisselmeier and Smith (1986) in an ideal adaptive regulation context.

4. Closed loop adaptive control system analysis

In this section we will first show that the adaptive control law (3.2)-(3.3a-h) ensures the estimated model admissibility, i.e. property P4. Then, the early stated control objective is achieved.

The following proposition will be useful in the remainder of the section.

Proposition 4.0 There exists a positive constant K_m , independent of β , such that for all positive t one has

$$m(t+1)/m(t) \leq K_m + \beta. \quad \nabla$$

The proof of this proposition follows *mutatis-mutandis* the proof of Proposition 2 in Giri *et al.* (1987).

The main step to ensure the estimated model admissibility property P4, consists in proving the following richness property.

Proposition 4.1. Consider the plant described by the model (2.1)-(2.2) in closed loop with the modified adaptive control law (3.2)-(3.3a-h) and suppose that the assumptions A1-A4 hold. Then there exist positive scalars λ and K_a (independent of μ) and a non-negative real sequence $\{\pi(t)\} \in S_a(K_a\mu)$ such that for any $t \in I$ and unit vector ω

$$\max_{4n+2 \leq t \leq 4n-2} |\omega^T \hat{\phi}(t+\tau)| \geq \lambda - \pi(t). \quad \nabla$$

Proof. See Appendix A.

The above proposition stipulates that the proposed adaptive control law provides the closed loop system with a self excitation capability, for sufficiently small μ .

Theorem 4.1. Consider the closed loop system of Proposition 4.1. There exists a positive scalar μ^* such that if $0 \leq \mu \leq \mu^*$ then:

- The plant estimated model is admissible, i.e. $\{\hat{\theta}(t)\}$ satisfies property P4
- There exists a positive scalar β^* such that if $0 < \beta \leq \beta^*$, then all the closed loop states are uniformly bounded for all initial states; namely the plant input and output
- In addition, the performance quantifiers

$$e_a(t) = P(\hat{\theta}_a(t))u(t) - A(\hat{\theta}_a(t))T(\hat{\theta}_a(t))v^*(t)$$

$$e_y(t) = P(\hat{\theta}_a(t))y(t) - B(\hat{\theta}_a(t))T(\hat{\theta}_a(t))v^*(t)$$

(with $P(\theta) = A(\theta)R(\theta) + B(\theta)S(\theta)$) are asymptotically zero in the ideal case, i.e. $\theta^*(t) = \theta^*(t-1) = \theta^*(0)$ and $\eta(t) = 0$. ∇

Part (a) means that even if $\hat{\theta}(t)$ becomes non-admissible for some t the intervals of non-admissibility are uniformly bounded.

The proof of part (a) is given in Appendix B. Parts (b) and (c) readily follow from part (a) as the sequence $\{\hat{\theta}_a(t)\}$ as defined in 3.3 satisfies properties P1-P4. Indeed, any linear control law combined with a parameter estimator satisfying some well defined properties (P_1 , P_2 , P_3 and P_4) provides a robust adaptive control. See for instance, Praly (1982), Samson (1983a,b), De Larminat (1986), De Larminat and Raynaud (1988), Giri *et al.* (1988b).

5. Conclusion

This paper has been written with two main objectives in mind. The first one is to show how to design a robustly stable indirect adaptive (linear) controller in a unified way, keeping in mind the available adaptive control theory, which is simply obtained by combining any suitable (satisfying A4) linear control design with a parameter estimator which is able to accommodate bounded disturbances, time varying parameters and unmodeled dynamics. In particular the less restrictive properties which determine the choice of the involved parameter estimator have been described, i.e. the robust stability properties P1-P4. The most crucial robust stability property, which is not necessarily satisfied by the available robust parameter estimators, is the estimated model admissibility with respect to the underlying linear control law.

The second objective is to provide a new solution to the estimated model admissibility problem, using an "active approach". The latter consists in ensuring that the signals involved in the adaptive closed loop system have a sufficient amount of excitation which makes it possible to get an admissible plant model. To this end, an *ad-hoc* modification of the control law is made. Such a modification consists of adding an internal impulse exciting signal while freezing the controller parameter, whenever the estimated model admissibility is lost. The proposed solution provides, in particular, new insights about the concept of persistent excitation in non-ideal situations.

It is however worth noticing that the robust stability in question means that the adaptive control system variables

remain uniformly bounded in spite of the considered non-idealities. A lot of work remains to be done about the performance of the adaptive controller in realistic situation.

References

- Anderson, B. D. O., R. R. Bitmead, C. R. Johnson, P. V. Kokotovic, R. L. Kosut, I. M. T. Mareels, I. Praly and B. D. Riedle (1986). *Stability of Adaptive Systems: Passivity and averaging analysis*. MIT Press, MA.
- Anderson, B. D. O. and R. M. Johnstone (1983). Adaptive systems and time-varying plants. *Int. J. Control*, **37**, 367–377.
- Anderson, B. D. O. and R. M. Johnstone (1985). Global adaptive pole positioning. *IEEE Trans. Aut. Control*, **AC-30**, 11–22.
- Åström, K. J. (1987). Adaptive feedback control. *IEEE Proc.*, **19**, 185–217.
- Åström, K. J., P. Hagander and J. Sternby (1984). Zeros of sampled systems. *Automatica*, **20**, 31–38.
- Bitmead, R. R., M. Gevers and V. Wertz (1990). *Adaptive Optimal Control: The Thinking Man's GPC*. Prentice-Hall International, Series in Systems and Control Engineering, Hemel Hempstead, UK.
- De Larminat, Ph. (1981). Unconditional stabilization of linear discrete systems via adaptive control. *Syst. Control Lett.*, **1**, 47–51.
- De Larminat, Ph. (1984). On the stabilizability condition in indirect adaptive control. *Automatica*, **20**, 793–798.
- De Larminat, Ph. (1986). Une solution robuste au problème de la stabilité dans la commande adaptative indirecte passive. *Commande Adaptative: Aspects Pratiques et Théoriques*. Édité par F. D. Landau et I. Dugard, Masson, Paris.
- De Larminat, Ph. and H. F. Raynaud (1988). A robust solution of the admissibility problem in indirect passive adaptive control without persistency of excitation. *Int. J. Adaptive Control Signal Proc.*, **2**, 98–100.
- Egardt, B. (1979). Stability of adaptive controllers. *Lecture Notes in Control and Information Sciences*, Vol. 20. Springer, Berlin.
- Egardt, B. and C. Samson (1982). Stable adaptive control of non-minimum phase systems. *Syst. Control Lett.*, **3**, 137–144.
- Elliot, H., R. Cristi and M. Das (1988). Global stability of adaptive pole placement algorithms. *IEEE Trans. Aut. Control*, **AC-30**, 348–356.
- Fuchs, J. J. (1980). Explicit self-tuning methods. *IEE Proc.*, **127**, Pt. D, 259–264.
- Fuchs, J. J. (1982). Sur la commande adaptative des systèmes linéaires discrets. Thèse d'Etat, Université de Rennes I.
- Giri, F., J. M. Dion, I. Dugard and M. M'Saad (1987). Robust pole placement direct adaptive control. *Proc. 26th CDC*, Los Angeles, U.S.A. and (1989) *IEEE Trans. Aut. Control*, **34**, 356–359.
- Giri, F., M. M'Saad, I. Dugard and J. M. Dion (1988a). Robust pole placement indirect adaptive controller. *Int. J. Adaptive Control Signal Proc.*, **2**, 33–47.
- Giri, F., M. M'Saad, J. M. Dion and I. Dugard (1988b). On the robustness of discrete-time indirect adaptive linear controllers. *Proc. 1988 IFAC Workshop On Robust Adaptive Control*, Newcastle, Australia, 1988, pp. 60–65.
- Goodwin, G. C. and K. S. Sin (1984). *Adaptive Prediction Filtering and Control*. Prentice-Hall, Englewood Cliffs, NJ.
- Goodwin, G. C., J. P. Norton and M. S. Viswanathan (1985). Persistency of excitation for non-minimal models of systems having purely deterministic disturbances. *IEEE Trans. Aut. Control*, **AC-30**, 589–592.
- Goodwin, G. C. and E. K. Teoh (1985). Persistency of excitation in the presence of possibly unbounded signals. *IEEE Trans. Aut. Control*, **AC-30**, 595–597.
- Ioannou, P. A. and P. V. Kokotovic (1983). Adaptive systems with reduced order models. *Lecture Notes in Control and Information Sciences*, Vol. 47. Springer, Berlin.
- Kreisselmeier, G. and K. S. Narendra (1982). Stable model reference control in the presence of bounded disturbances. *IEEE Trans. Aut. Control*, **AC-27**, 1169–1175.
- Kreisselmeier, G. (1985). An approach to stable indirect adaptive control. *Automatica*, **21**, 425–431.
- Kreisselmeier, G. (1986a). A robust indirect adaptive control approach. *Int. J. Control*, **43**, 161–175.
- Kreisselmeier, G. (1986b). Adaptive control of a class of slowly time-varying plants. *Syst. Control Lett.*, **8**, 97–103.
- Kreisselmeier, G. and M. Smith (1986). Stable adaptive regulation of arbitrary n th-order plants. *IEEE Trans. Aut. Control*, **AC-31**, 299–305.
- Lozano, R. E. and G. C. Goodwin (1985). A globally convergent adaptive pole placement algorithm without a persistency of excitation requirement. *IEEE Trans. Aut. Control*, **AC-30**, 798–798.
- Middleton, R. H., G. C. Goodwin, D. J. Hill and D. O. Mayne (1988). Design issues in adaptive control. *IEEE Trans. Aut. Control*, **AC-33**, 80–88.
- Middleton, R. H. and G. C. Goodwin (1988). Adaptive control of time-varying linear systems. *IEEE Trans. Aut. Control*, **AC-33**, 150–155.
- M'Saad, M., R. Ortega and F. D. Landau (1985). Adaptive controllers for discrete time systems with arbitrary zeros: An overview. *Automatica*, **21**, 413–423.
- Narendra, K. S. and A. M. Annaswamy (1987). Persistent excitation in adaptive systems. *Int. J. Control*, **45**, 127–160.
- Peterson, B. B. and K. S. Narendra (1982). Bounded error adaptive control. *IEEE Trans. Aut. Control*, **AC-27**, 1161–1168.
- Polderman, J. W. (1989). A state space approach to the problem of adaptive pole assignment. *Math. Control Signals Syst.*, **2**, 71–94.
- Praly, L. (1982). MIMO indirect adaptive control: stability and robustness. Technical Report, December 1982, C.A.T. Ecole des Mines, Fontainebleau, France.
- Praly, L. (1983). Robustness of indirect adaptive control based on pole placement design. *Proc. 1st IFAC Workshop on Adaptive Systems in Control and Signal Processing*, San Francisco, CA.
- Rohrs, C. E., V. Balas, M. Athans and G. Stein (1985). Robustness of adaptive control algorithms in the presence of unmodeled dynamics. *IEEE Trans. Aut. Control*, **AC-30**, 881–889.
- Samson, C. (1982). An adaptive IQ controller for non-minimum phase systems. *Int. J. Control*, **3**, 389–397.
- Samson, C. (1983a). Stability of adaptively controlled systems subject to bounded disturbances. *Automatica*, **19**, 81–86.
- Samson, C. (1983b). Problèmes en identification et commande des systèmes dynamiques. Thèse d'Etat, Université de Rennes I.
- Isakakis, K. and P. A. Ioannou (1987). Adaptive control of linear time-varying plants: New controller structure. *Proc. Aut. Control Conf.*, pp. 583–588.

Appendix A. Proof of Proposition 4.1

The proof given in this appendix is adapted from Kreisselmeier and Smith (1986). Consider the plant model (2.1)–(2.2) together with the adaptive control law (3.2)–(3.3a, b); one has for all t

$$(R(\hat{\theta}_a(t))A(\hat{\theta}^*(t)) + S(\hat{\theta}_a(t))B(\hat{\theta}^*(t)))x(t) \\ + T(\hat{\theta}_a(t))x^*(t) - S(\hat{\theta}_a(t))\eta(t) + \alpha(t) + \beta(t) = (A-1a)$$

Defining the state vector $x(t)$ as

$$x^T(t) = [x \ q^{-1} \ q^{-2} \ \dots \ q^{-(n-1)}]^T, \quad (A-1b)$$

it follows that

$$x(t+1) = H(t)x(t) + h[tT(\hat{\theta}_a(t))x^*(t) \\ - S(\hat{\theta}_a(t))\eta(t) + \alpha(t) + \beta(t)] \quad (A-2)$$

with

$$H(t) = \begin{bmatrix} \gamma(t) & \gamma(t) & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & \dots & 0 \end{bmatrix}, \quad h = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix} \quad (A-3)$$

where $\{\gamma_i(t)\}_{i=0, \dots, 2n-1}$ are the coefficients of the polynomial $R(\hat{\theta}_a(t))A(\theta^*(t)) + S(\hat{\theta}_a(t))B(\theta^*(t))$. Let k be an integer such that $-2n \leq k \leq -1$ and $f(t) = \alpha(t) + \beta(t)$; the equation (A.2) may be rewritten as

$$\begin{aligned} x(t+k+1) &= H(t)x(t+k) + h[T(\hat{\theta}_a(t+k))y^*(t+k) \\ &\quad + S(\hat{\theta}_a(t+k))\eta(t+k) + f(t+k)] \\ &\quad + (H(t+k) - H(t))x(t+k). \end{aligned} \quad (\text{A.4})$$

This is, for all $t \geq 0$

$$\begin{aligned} x(t+k+i+1) &= H(t)^{i+1}x(t+k) + \sum_{j=0}^i H(t)^{i-j}h \\ &\quad \times [T(\hat{\theta}_a(t+k+j))y^*(t+k+j) \\ &\quad + S(\hat{\theta}_a(t+k+j))\eta(t+k+j) + f(t+k+j)] \\ &\quad + \sum_{j=0}^i H(t)^{i-j}(H(t+k+j) - H(t))x(t+k+j). \end{aligned} \quad (\text{A.5})$$

As the characteristic equation of the system (A.2)–(A.3) is given by

$$\lambda^{2n} + \gamma_1(t)\lambda^{2n-1} + \dots + \gamma_{2n-1}(t)\lambda = 0$$

it follows that multiplying (A.5) by $\gamma_{2n-i-1}(t)$ and using the Cayley–Hamilton theorem leads to

$$\begin{aligned} \sum_{i=0}^{2n-1} \gamma_{2n-i-1}(t)x(t+k+i+1) &= \mathcal{K}(t)[\Phi(t+k+2n-1) \\ &\quad + Y_T^*(t+k+2n-1) - N_v(t+k+2n-1) + \Gamma(t+k)] \end{aligned} \quad (\text{A.6})$$

with

$$\gamma_0(t) = 1$$

$$\mathcal{K}(t) = [h, H(t)h, \dots, H(t)^{2n-1}h]$$

$$[\gamma_{2n-1}(t), \dots, \gamma_1(t), 1]$$

$$\begin{bmatrix} \gamma_1(t) \\ \vdots \\ 1 \end{bmatrix} = 0 \quad (\text{A.7})$$

$$\Phi(t+k+2n-1) = [f(t+k) \dots f(t+k+2n-1)]^T \quad (\text{A.8})$$

$$\begin{aligned} Y_T^*(t+k+2n-1) &= [T(\hat{\theta}_a(t+k))y^*(t+k) \\ &\quad \dots T(\hat{\theta}_a(t+k+2n-1))y^*(t+k+2n-1)]^T \end{aligned} \quad (\text{A.9})$$

$$\begin{aligned} N_v(t+k+2n-1) &= [S(\hat{\theta}_a(t+k))\eta(t+k) \\ &\quad \dots S(\hat{\theta}_a(t+k+2n-1))\eta(t+k+2n-1)]^T. \end{aligned} \quad (\text{A.10})$$

$$\begin{aligned} \Gamma(t+k) &= \sum_{i=0}^{2n-1} \gamma_{2n-i-1}(t) \sum_{j=0}^i H(t)^{i-j} \\ &\quad \times (H(t+k+j) - H(t))x(t+k+j). \end{aligned} \quad (\text{A.11})$$

Substituting $t+2n-1$ to t in (A.6) yields

$$\begin{aligned} \sum_{i=0}^{2n-1} \gamma_{2n-i-1}(t+2n-1)x(t+k+i+2n) \\ &= \mathcal{K}(t+2n-1)[\Phi(t+k+4n-2) + Y_T^*(t+k+4n-2) \\ &\quad + N_v(t+k+4n-2)] + \Gamma(t+k+2n-1) \end{aligned} \quad (\text{A.12})$$

As the plant model and the controller parameters are uniformly bounded, there exists a positive scalar K_v such that for any $t \in [0, 2n-1]$ and any $\tau \in [0, 4n-2]$, one has:

$$|\gamma_i(t+\tau)| \leq K_v \quad (\text{A.13})$$

Premultiplying (A.12) by an arbitrary unit vector v yields

$$\begin{aligned} \left| \sum_{i=0}^{2n-1} \gamma_{2n-i-1}(t+2n-1)v^T x(t+k+i+2n) \right| \\ &\leq |v^T \mathcal{K}(t+2n-1)\Phi(t+k+4n-2)| \\ &\quad + |v^T \mathcal{K}(t+2n-1)Y_T^*(t+k+4n-2)| \\ &\quad + |v^T \mathcal{K}(t+2n-1)N_v(t+k+4n-2)| \\ &\quad + |v^T \Gamma(t+k+2n-1)| \end{aligned} \quad (\text{A.14})$$

Since $-2n \leq k \leq -1$ and using (A.13) it follows that

$$\begin{aligned} \left| \sum_{i=0}^{2n-1} \gamma_{2n-i-1}(t+2n-1)v^T x(t+k+i+2n) \right| \\ \leq 2nK_v \max_{0 \leq \tau \leq 4n-2} |v^T x(t+\tau)|. \end{aligned} \quad (\text{A.15})$$

Furthermore, it follows from (A.3) and (A.7) that $\mathcal{K}(t)$ is a matrix with all singular values equal to 1. That is, for all $t \in I$ one has from (3.3a–h)

$$f(t+\tau) = 0 \quad \text{for } \tau \in \{0, 1, \dots, 2n-2, 2n, \dots, 4n-2\}$$

and hence (A.8) can be rewritten as

$$\Phi(t+k+4n-2)^T = [0 \dots 0 \ 1 \ 0 \dots 0]f(t+2n-1)$$

where the 1 is at the $(1-k)$ th entry. Thus there exists at least one $k \in [-2n, -1]$ depending on the choice of v such that

$$|v^T \mathcal{K}(t+2n-1)\Phi(t+k+4n-2)| \geq \frac{f(t+2n-1)}{2n} \quad (\text{A.16})$$

Moreover, taking into account the uniform boundedness of the $\{y^*(t)\}$ and $\{\hat{\theta}_a(t)\}$ sequences as well as (A.9) and (3.3a–h), one has

$$|v^T \mathcal{K}(t+2n-1)Y_T^*(t+k+4n-2)| \leq \frac{\alpha(t+2n-1)}{\sqrt{2n}} \quad (\text{A.17})$$

And from (A.10), (A.11) and (A.13), it follows that there exist positive scalars K_v and K_f independent of μ and v such that

$$|v^T \mathcal{K}(t+2n-1)N_v(t+k+4n-2)| \leq K_v \max_{4n+2n-1 \leq \tau \leq 4n-2} |\eta(t+\tau)| \quad (\text{A.18a})$$

$$|v^T \Gamma(t+2n-1)| \leq K_f \max_{4n+2n-1 \leq \tau \leq 4n-2} \|\Gamma(t+\tau)\| \quad (\text{A.18b})$$

Combining (A.14)–(A.18) leads to

$$\begin{aligned} 2nK_v \max_{4n+2n-1 \leq \tau \leq 4n-2} |v^T x(t+\tau)| &\geq \frac{\beta m(t+2n-1)}{2n} \\ &\quad - K_v \max_{4n+2n-1 \leq \tau \leq 4n-2} |\eta(t+\tau)| - K_f \max_{4n+2n-1 \leq \tau \leq 4n-2} \|\Gamma(t+\tau)\| \end{aligned} \quad (\text{A.19})$$

On the other hand, the measurement vector $\phi(t)$ in Assumption A3 may be expressed in terms of the partial state $z(t)$ in (2.1) as follows

$$\begin{aligned} \phi(t)^T &= [-B(\theta^*(t-1)) \dots q^{-n+1}B(\theta^*(t-1)), \\ &\quad A(\theta^*(t-1)) \dots q^{-n+1}A(\theta^*(t-1))]z(t-1) \\ &\quad + [-1 - q^{-1} \dots - q^{-n+1} \ 0 \dots 0]\eta(t-1). \end{aligned} \quad (\text{A.20})$$

In order to link the vector $\phi(t)$ to $x(t)$, let us define the following vector

$$\begin{aligned} \phi'(t)^T &= [-B(\theta^*(t-1)), \dots, -q^{-n+1}B(\theta^*(t-1)), \\ &\quad A(\theta^*(t-1)), \dots, q^{-n+1}A(\theta^*(t-1))]z(t-1) \end{aligned} \quad (\text{A.21})$$

where \cdot denotes the usual polynomial multiplication. Furthermore letting $N(t) = \phi(t) - \phi'(t)$, it follows from Assumption A3 and (2.6) that there exists a positive scalar K_{N1} independent of μ such that

$$\{\|\hat{N}(t)\|\} \in S_n(K_{N1}\mu) \quad (\text{A.22})$$

In view of Assumption A2, the $(2n)$ polynomials appearing in the right-hand side of (A.21) are linearly independent over the reals. This together with (A.1b) leads to

$$\phi'(t) = \Lambda(t)x(t)$$

where $\Lambda(t)$ is a full rank $2n \times 2n$ dimension matrix satisfying $\lambda_{\min}(\Lambda(t)) \geq \sigma_{\min} > 0$ for all $t \geq 0$; λ_{\min} being the minimal singular value of $\Lambda(t)$. This yields

$$\phi(t) = \Lambda(t)x(t) + N(t) \quad (\text{A.23})$$

This is, premultiplying (A.23) by a unit vector w leads to

$$|w^T \phi(t)| \geq |w^T \Lambda(t)x(t)| - \|N(t)\| \quad (\text{A } 24)$$

Letting $w(t)^T = w^T \Lambda(t) / \|w^T \Lambda(t)\|$, it follows from (A.24) that

$$\begin{aligned} \max_{4n+2 \leq t \leq 4n-2} |w^T \phi(t+\tau)| \\ \geq \sigma_{\min} \max_{4n+2 \leq t \leq 4n-2} |w(t+\tau)^T x(t+\tau)| \\ - \max_{4n+2 \leq t \leq 4n-2} \|N(t+\tau)\| \end{aligned} \quad (\text{A } 25)$$

Using (A.19) with $v = w(t+\tau)$, (A.25) yields for any $t \in I$

$$\begin{aligned} \max_{4n+2 \leq t \leq 4n-2} |w^T \phi(t+\tau)| &\geq \left(\frac{\beta \sigma_{\min}}{4n^2 K_y} \right) m(t+2n-1) \\ &- \left(\frac{K_1 \sigma_{\min}}{2n K_y} \right) \max_{4n+2 \leq t \leq 4n-2} \|\eta(t+\tau)\| \\ &- \left(\frac{K_1 \sigma_{\min}}{2n K_y} \right) \max_{4n+2 \leq t \leq 4n-2} \|\Gamma(t+\tau)\| \\ &- \max_{4n+2 \leq t \leq 4n-2} \|N(t+\tau)\| \end{aligned} \quad (\text{A } 26)$$

In other respects, dividing each member of the above inequality by $m(t+4n-1)$ and using Proposition 4.0, it follows that there exists a positive constant K_1 (independent of μ) such that:

$$\max_{4n+2 \leq t \leq 4n-2} |w^T \phi(t+\tau)| \geq \frac{K_1 \beta \sigma_{\min}}{4n^2 K_y} \pi(t) \quad (\text{A } 27)$$

where

$$\begin{aligned} \pi(t) = K_2 \left[\left(\frac{K_1 \sigma_{\min}}{2n K_y} \right) \max_{4n+2 \leq t \leq 4n-2} \|\eta(t+\tau)\| \right. \\ \left. + \left(\frac{K_1 \sigma_{\min}}{2n K_y} \right) \max_{4n+2 \leq t \leq 4n-2} \|\Gamma(t+\tau)\| \right. \\ \left. + \max_{4n+2 \leq t \leq 4n-2} \|N(t+\tau)\| \right] \end{aligned} \quad (\text{A } 28)$$

for some positive scalar K_2 (independent of μ).

Notice that since $\{\|N(t)\|\}$, $\{\|\Gamma(t)\|\}$ and $\{\|\eta(t)\|\}$ are $K\mu$ -asymptotically small in the mean, for some $K > 0$, their maximum values on any finite interval are $K'\mu$ asymptotically small in the mean, for some $K' > 0$, depending on K .

These being, taking into account (A.22) and Assumption A3 and properties P1 and P3, there exists a positive constant K_3 such that:

$$\{\pi(t)\} \in S_a(K_3 \mu) \quad (\text{A } 29)$$

which establishes Proposition 4.1 with $\lambda = K_1 \beta \sigma_{\min} / 4n^2 K_y$ □□□

Appendix B: Proof of Theorem 4.1 (part a)

The proof will be made by contradiction. To this end, let $\{t_k\}$ be the time sequence such that for all $k \in \mathbb{N}$, $\hat{\theta}(t_k) \in \mathcal{D}_\mu$ and assume that the property P4 does not hold. Then the incremental sequence $\{\tau_k = t_k - t_{k-1}\}$ is not uniformly bounded, and hence there exists a non-decreasing subsequence $\{\tau_{k_j}\}$ such that $\tau_{k_j} \rightarrow \infty$ as $j \rightarrow \infty$.

This is, one has

$$\hat{\theta}(t) \notin \mathcal{D}_\mu \text{ for any positive integer } t \in [t_{(k_j-1)}, t_{k_j}]$$

And from Assumption A4, it follows that there exists a positive scalar $r > \varepsilon$ such that for any $t \in [t_{(k_j-1)}, t_{k_j}]$, one has $\|\hat{\theta}(t)\| = \|\hat{\theta}(t) - \theta^*(t)\| \geq r$, which implies

$$\inf_{t_{(k_j-1)} \leq t \leq t_{k_j}} \|\hat{\theta}(t)\| \geq r \quad (\text{B } 1)$$

For any fixed $t \in [t_{(k_j-1)}, t_{k_j}]$, applying Proposition 4.1 for $w = \hat{\theta}(t) / \|\hat{\theta}(t)\|$ it follows that for a sufficiently large j

$$\max_{t \in [t_{(k_j-1)}, t_{k_j}]} \left| \phi(t+\tau)^T \frac{\hat{\theta}(t)}{\|\hat{\theta}(t)\|} \right| \geq \lambda - \pi(t)$$

where $\tau_{\max} = 4n-2$ and

$$\{\pi(t)\} \in S_a(K_3 \mu) \quad (\text{B } 2)$$

This leads to

$$\|\hat{\theta}(t)\| \leq \frac{1}{\lambda} \max_{t_{(k_j-1)} \leq t \leq t_{k_j}} |\phi(t+\tau)^T \hat{\theta}(t)| + \frac{K_3 \pi(t)}{\lambda} \quad (\text{B } 3)$$

On the other hand, for any t and any integer $t \in [-t_{\max}, t_{\max}]$ one has

$$\begin{aligned} |\phi(t+\tau)^T \hat{\theta}(t)| &\leq |\phi(t+\tau)^T \theta(t+\tau)| \\ &+ |\phi(t+\tau)^T (\hat{\theta}(t+\tau) - \theta(t+\tau))| \\ &\leq \sum_{i=1}^{t_{\max}} |\phi(t+\tau)^T \theta(t+\tau)| \\ &+ \sum_{i=1}^{t_{\max}} \|\phi(t+\tau)\| \|\hat{\theta}(t+\tau) - \theta(t+\tau)\| \end{aligned}$$

But from the plant model (2.1)–(2.2), one gets

$$\phi(t+\tau)^T \theta(t+\tau) = \gamma(t+\tau) + \eta(t+\tau)$$

where

$$\gamma(t) = y(t) - \theta(t)^T \phi(t) \quad \text{and} \quad \eta(t) = A(\theta^*(t)) \eta(t)$$

Letting

$$\begin{aligned} \Delta(t) &= \sum_{i=1}^{t_{\max}} |\gamma(t+i)| + \sum_{i=1}^{t_{\max}} \|\eta(t+i)\| \\ &+ \sum_{i=1}^{t_{\max}} \|\hat{\theta}(t+i) - \theta(t+i)\| \end{aligned}$$

one has

$$\max_{t \in [t_{(k_j-1)}, t_{k_j}]} |\phi(t+\tau)^T \hat{\theta}(t)| \leq \Delta(t)$$

This together with (B.3) yields for any $t \in [t_{(k_j-1)}, t_{k_j}]$

$$\|\hat{\theta}(t)\| \leq \frac{\Delta(t) + K_3 \pi(t)}{\lambda} \quad (\text{B } 4)$$

On the other hand, it follows from Assumptions A1 and A3 and property P2 that there exists a positive constant K_4 (independent of μ) such that

$$\{\Delta(t)\} \in S_a(K_4 \mu) \quad (\text{B } 5)$$

Noting that $t_{k_j} - t_{k_j-1} = t_{(k_j-1)}$, it therefore follows from (B4) that

$$\begin{aligned} \inf_{t_{(k_j-1)} \leq t \leq t_{k_j}} \|\hat{\theta}(t)\| &\leq \frac{1}{\tau_{k_j} - 1} \sum_{i=1}^{\tau_{k_j}-1} \|\hat{\theta}(t+i)\| \\ &\leq \frac{1}{\tau_{k_j} - 1} \sum_{i=1}^{\tau_{k_j}-1} \frac{\Delta(t+i) + K_3 \pi(t+i)}{\lambda} \end{aligned} \quad (\text{B } 6)$$

Since $\{t_{(k_j-1)}\}$ and $\{\tau_{k_j}\} \rightarrow \infty$ as $j \rightarrow \infty$, it follows from (B.6), using (B.2) and (B.5) that, for j sufficiently large

$$\inf_{t_{(k_j-1)} \leq t \leq t_{k_j}} \|\hat{\theta}(t)\| \leq \frac{(K_4 + K_3 K_4) \mu}{\tau_{k_j}} \quad (\text{B } 7)$$

Thus, if μ^* is such that $0 < \mu^* < \lambda r / 2(K_4 + K_3 K_4)$, then the right-hand side of (B.7) is smaller than $r/2$, which contradicts (B.1). Therefore, no divergent subsequence $\{\tau_{k_j}\}$ can be extracted from $\{\tau_k\}$. Hence, $\{\tau_k\}$ is uniformly bounded. This establishes the part (a) of Theorem 4.1 □□□

Brief Paper

Quadratic Stabilizability of Uncertain Systems: A Two Level Optimization Setup*

KEQIN GU,^{†‡} Y. H. CHEN,[‡] M. A. ZOHDY[§] and NAN K. LOH[§]

Key Words—Robust control, stability, control system design, convex programming, optimization

Abstract—The problem of stabilizing linear systems subject to possibly fast time varying uncertainties is investigated. Necessary and sufficient conditions of quadratic stabilizability are discussed. The design process is formulated as a two level optimization process, which can be simplified if the uncertainty is bounded by a hyperpolyhedron.

1 Introduction

THE PROBLEM of stability and feedback control of uncertain linear systems has drawn much attention recently. Numerous criteria have been devised to characterize the uncertainties, either constant or time varying, such that the stability of the system is guaranteed if the criteria are satisfied (Eslami and Russell, 1980; Yedavalli, 1985a, b, 1988; Patel *et al.*, 1977; Patel and Tock, 1980; Zhou and Khargonekar, 1987). However, these criteria are generally rather conservative, and some impose very restrictive assumptions. Barmish (1985), by an elegant mathematical maneuver, derived a necessary and sufficient condition of stabilizability of a fairly general class of uncertainties. However, the stabilizability condition is difficult to check, and the stabilizing controller design is tedious. Petersen, Khargonekar, and Zhou published a number of papers on the stabilizability of uncertain systems of the following form (Petersen, 1987, 1988; Khargonekar *et al.*, 1988):

$$\dot{x} = (A + \Delta A(t))x + (B + \Delta B(t))u,$$

where

$$(\Delta A, \Delta B) = D F(t) (E_1, E_2), \quad \text{and} \quad F^T(t) F(t) \leq I$$

Many ingenious results have been derived. A necessary and sufficient condition of stabilizability of the above system has recently been derived, and the expression of stabilizing control has been obtained (Khargonekar *et al.*, 1988). While many uncertainties can be expressed as a subset of the above uncertainty expression by the "over bounding" technique, the resulting criterion is often extremely conservative. Gu *et al.* (1989) proposed a necessary and sufficient condition of quadratic stability, which can be checked by an explicit algorithm when the uncertainty lies in a hyperpolyhedron.

This paper extends the result of Gu *et al.* (1989) to linear

feedback control of uncertain systems. The major contribution of this paper is necessary and sufficient conditions of quadratic stabilizability of arbitrary uncertainty by linear feedback control are reformulated to a form of two level optimization process, where the inner level optimum can be reached by the vertices if the uncertainty set is a hyperpolyhedron, and any local maximum of the outer level is also the global maximum.

2 Preliminary notes

For the convenience of development, a number of notations are introduced. All the matrices considered in this paper are real. The set of $n \times m$ matrices will be denoted as $R^{n \times m}$, and the set of column vectors with n components is R^n . The set of symmetric $n \times n$ matrices will be denoted as S^n . S^n_+ is the set of positive definite matrices. The closure S^n_+ is clearly the set of positive semi-definite matrices. For a scalar c , S^n_c is defined as

$$S^n_c = \{Q | Q = cI + S^n_+\}$$

It is important to notice that S^n_c is a convex cone with the apex at cI , i.e. for any $Q_1, Q_2 \in S^n_c$ and $\alpha \in [0, 1]$, then $\beta Q_1 + (1 - \beta)Q_2 \in S^n_c$, and $cI + \alpha(Q_1 - cI) \in S^n_c$. Also notice that for any compact set, $C \in S^n_+$ implies the existence of an $\epsilon > 0$ such that $C \in S^n_\epsilon$. Analogously, S^n_- is the set of all negative definite matrices, and

$$S^n_- = \{Q | Q = -cI + S^n_+\}$$

Let Σ be an arbitrary set in a vector space, then $\text{conv}(\Sigma)$ denotes its convex hull, i.e.

$$\text{conv}(\Sigma) = \{C_1 + \dots + C_n | C_i \in \Sigma, \sum_{i=1}^n \alpha_i = 1, \alpha_i \geq 0\}$$

Let C be a convex set, then $\text{protr}(C)$ denotes the set of protruded points of C . For $D \in R^{n \times n}$, $G \in R^{n \times n}$, a linear operator $T^D_C : S^n_+ \rightarrow S^n_+$ is defined as

$$T^D_C P = W, \quad \text{if} \quad D^T(GP + PG^T)D = W$$

T^D_C is abbreviated as T_C . It is important to note that T^D_C is itself a linear map

$$T^D_C(\alpha_1 P_1 + \alpha_2 P_2) = \alpha_1 T^D_C P_1 + \alpha_2 T^D_C P_2, \quad \text{for} \quad \alpha_1, \alpha_2 \in R$$

The uncertain systems considered in this paper can be expressed as

$$\dot{x} = (A + \Delta A(t))x + (B + \Delta B(t))u, \quad (1)$$

where $x \in R^n$ is the vector of state variables, $u \in R^m$ is the vector of control input, $A \in R^{n \times n}$ is the nominal system matrix, and $B \in R^{n \times m}$ is the nominal input matrix. The uncertainty is represented by $(\Delta A(t), \Delta B(t))$, which is an arbitrary Lebesgue measurable matrix function satisfying

$$(\Delta A(t), \Delta B(t)) \in \Omega \quad \text{for all } t \geq 0$$

‡ A protruded point p of a set C is such a point, for any linear segment which contains p and is contained in C , p is one of its end points (Gu *et al.*, 1989).

* Received 23 May 1989; revised 19 February 1990; received in final form 9 April 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor P. Dorato under the direction of Editor H. Kwakernaak.

† Center for Robotics and Advanced Automation, School of Engineering and Computer Science, Oakland University, Rochester, MI 48309, U.S.A.

‡ The George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A. Author to whom all correspondence should be addressed.

§ Currently at the Department of Mechanical Engineering, Southern Illinois University at Edwardsville, Edwardsville, IL 62026, U.S.A.

where Ω is a compact set in the finite dimensional vector space $R^{n \times (n+m)}$.

Remark 1. Typically, Ω is contained in a subspace of $R^{n \times (n+m)}$. In this case, it is often beneficial to express (1) as

$$\dot{x} = \left(A + \sum_{i=1}^m q_i(t) A_i \right) x + \left(B + \sum_{i=1}^m q_i(t) B_i \right) u, \quad (2)$$

where $((A_i, B_i), i = 1, \dots, m)$ is a set of basis vectors (they are actually matrices) of the subspace containing Ω , and

$$q(t) = (q_1(t), q_2(t), \dots, q_m(t))^T$$

is a Lebesgue measurable vector function, satisfying

$$q(t) \in \Omega_q \quad \text{for all } t \geq 0$$

Corresponding to the compactness of Ω , Ω_q is also a compact set in R^m . In applications, equation (2) is often directly available from physical model, with $q_i(t)$ being the uncertain factors. Since $R^{n \times (n+m)}$ is its own subspace, (1) and (2) are equivalent expressions. Although the $q_i(t)$ s should be linearly independent, they can be functionally dependent. An example is

$$\Delta A(t) = \begin{pmatrix} p(t) & p^2(t) \\ p^2(t) & p(t) \end{pmatrix} \quad |p(t)| \leq 1,$$

which can be written as

$$\Delta A(t) = q_1(t) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + q_2(t) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

with

$$\Omega_q = \{(p, p^2) : |p| \leq 1\}.$$

The essential problem is to stabilize system (1) by a linear feedback. We will restrict our attention to "quadratic stability", which was proposed by Harmish (1985):

Definition 1. A linear feedback control $u = -Kx$ is said to quadratically stabilize system (1) if there exists a matrix $P \in S_n^+$ and a scalar $\alpha > 0$ such that

$$x^T [P(A + E) - (B + F)K] + (A + E - (B + F)K)^T P]x \leq -\alpha \|x\|^2 \quad (3)$$

holds for arbitrary $x \in R^n$ and $(E, F) \in \Omega$. A system of form (1) is said to be quadratically stabilizable by a linear feedback control if there exists a linear feedback law $u = -Kx$ which quadratically stabilizes it.

3. Systems without uncertainties in B

As a first step of development, it is assumed in this section that no uncertainties exist in the input matrix

Assumption 1. The uncertain system considered in this section can be written as

$$\dot{x} = (A + \Delta A(t))x + Bu, \quad (4)$$

where $\Delta A(\cdot)$ is an arbitrary Lebesgue measurable matrix function satisfying

$$\Delta A(t) \in \Omega \quad \text{for all } t \geq 0, \quad (5)$$

where Ω is a compact convex set in $R^{n \times n}$, $B \in R^{n \times r}$ has full rank, and $n > r$.

The stabilizability conditions of such a system are stated in the following theorem, which is rephrased from a theorem by Hollot and Harmish (1980):

Theorem 1. Let $D \in R^{(n-r) \times (n-r)}$, such that

$$V = \begin{pmatrix} B^T \\ D^T \end{pmatrix} \quad (6)$$

is nonsingular, and let its inverse be partitioned as

$$V^{-1} = \begin{pmatrix} \hat{B} & \hat{D} \end{pmatrix}, \quad (7)$$

where $\hat{B} \in R^{n \times r}$ and $\hat{D} \in R^{(n-r) \times (n-r)}$. Then system (1) satisfying

Assumption 1 is quadratically stabilizable by a linear feedback control if and only if there exists an $S \in S_n^+$ such that

$$T_{A,E}^D S = -\hat{D}^T [(A + E)S + S(A + E)^T] \hat{D} \in S_n^+, \quad \text{for all } E \in \Omega. \quad (8)$$

Remark 2. If the condition of the theorem is satisfied, then a stabilizing controller can be expressed as

$$K = \frac{1}{\epsilon} R^{-1} B^T P,$$

where $P = S^{-1}$, $R \in S_r^+$, and $\epsilon > 0$ sufficiently small such that

$$-T_{A,E}^D P^{-1} - \frac{2}{\epsilon} R^{-1} + [\hat{B}^T (T_{A,E} P^{-1}) \hat{D}] \times [T_{A,E}^D P^{-1}]^{-1} [\hat{B}^T (T_{A,E} P^{-1}) \hat{D}]^T \in S_r^+ \quad \text{for all } E \in \Omega. \quad (9)$$

For practical applications, the above theorem needs to be put in a form which is more easily checked, and such that it is easier to design the feedback controller if the system is indeed quadratically stabilizable. Theorems 2 and 3 serve this purpose. The following lemma is instrumental for further development:

Lemma 1. If $U, V, W \in S_n^+$, $0 < \alpha < 1$, and $W = \alpha U + (1 - \alpha)V$. Then

$$\lambda_{\min}(W) \geq \alpha \lambda_{\min}(U) + (1 - \alpha) \lambda_{\min}(V) \quad (10)$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of the designated matrix.

Proof.

$$\begin{aligned} \lambda_{\min}(W) &= \min_{\|x\|=1} x^T W x \\ &= \min_{\|x\|=1} x^T [\alpha U + (1 - \alpha)V] x \\ &= \min_{\|x\|=1} [\alpha x^T U x + (1 - \alpha)x^T V x] \\ &\geq \alpha \min_{\|x\|=1} x^T U x + (1 - \alpha) \min_{\|x\|=1} x^T V x \\ &= \alpha \lambda_{\min}(U) + (1 - \alpha) \lambda_{\min}(V). \end{aligned}$$

An immediate consequence of this lemma is that

$$\lambda_{\min}(W) \geq \min \{ \lambda_{\min}(U), \lambda_{\min}(V) \} \quad (11)$$

The above lemma and the linearity of the operator T^D with respect to the subscripts lead to the following theorem:

Theorem 2. Let $S \in S_n^+$, $\hat{D} \in R^{(n-r) \times (n-r)}$, $A \in R^{n \times n}$ and Ω convex. Let

$$\gamma = \min_{E \in \Omega} \lambda_{\min}[T_{A,E}^D(S)],$$

then γ can be reached by a protruded point of Ω .

Proof. Since Ω is compact, the minimum γ can be reached by some point $E_0 \in \Omega$. Let H_0 be an affine subspace of $R^{n \times n}$ of minimum dimension such that $\Omega \subset H_0$. Also let $\Omega_0 = \Omega$. If E_0 is a protruded point, then the proof is complete. Otherwise, there exists a line segment L , such that $E_0 \in L$, $L \subset \Omega_0$, and both ends of the line segment are on the boundary of Ω_0 . Since $T_{A,E}^D S$ is linear with respect to E , according to (11), the minimum can be reached by one of the end points of the line segment. Let this minimum end point be E_1 . If E_1 is a protruded point, then the proof is again completed. Otherwise, since Ω_0 is convex, and E_1 is its boundary point, there exists a hyperplane H_1 of H_0 supporting Ω_0 at E_1 (i.e. Ω_0 is in one side of H_1 and $E_1 \in H_1$) (Luenberger, 1969). Then it is easily seen that all the line segments of Ω_0 which contain E_1 but do not end at E_1 are contained in H_1 , and H_1 is one dimension less than H_0 . Let $\Omega_1 = H_1 \cap \Omega_0$, then Ω_1 is also convex, and all the protruded points of Ω_1 are also protruded points of Ω_0 . Repeat the same process on the triple (Ω_1, H_1, E_1) , either a protruded point is found or a new triple (Ω_2, H_2, E_2) can be produced with H_2 one dimension less than H_1 . Therefore this process is

terminating, and a protruded point can be found, which is a minimum point.

Theorem 2 allows us to search only the protruded points for γ . When Ω is a hyperpolyhedron, there are only finite number of protruded points (i.e. the vertices). Another point to be noticed is that Theorem 1 is still valid if " $S \in S^*$ " is replaced by " $S \in S_1^*$ and $\|S\| = 1$ ", where $\|\cdot\|$ is any norm of matrices. A convenient choice in this case is the sum of singular value norm, which reduces to the trace for positive semi-definite matrices, and therefore has the nice property of being a linear function of the matrix

$$\|S\| = \text{tr}(S), \text{ if } S \in S_1^*$$

This selection of norm results in the following theorem, which is important for numerical computation

Theorem 3 Let

$$\phi(S) = \min_{T \in U} \lambda_{\min}(T_{A+T}^P S)$$

and

$$\psi_2(S) = \min\{\phi(S), \xi \lambda_{\min}(S)\}$$

where $\xi > 0$. Also let

$$W = \{S \in S_1^* : \text{tr}(S) = 1\}$$

Then any local maximum of $\psi_2(S)$ in W is also the global maximum in W

Proof. Since the trace operator is linear and S_1^* is convex, W is also convex. The proof is proceeded by contradiction. Assume that S_1 is a local maximum point of $\psi_2(S)$, and S_2 is the global maximum point, $\psi_2(S_1) < \psi_2(S_2)$. Let

$$\omega(E, S) = \lambda_{\min}(T_{A+T}^P S)$$

Since W is convex,

$$\alpha S_1 + (1 - \alpha) S_2 \in W \text{ for all } 0 < \alpha < 1$$

By Lemma 1 and the linearity of T_{A+T}^P , it is easily seen

$$\omega(E, \alpha S_1 + (1 - \alpha) S_2) = \alpha \omega(E, S_1) + (1 - \alpha) \omega(E, S_2)$$

Take the minimum of both sides

$$\begin{aligned} \min_{T \in U} \omega(E, \alpha S_1 + (1 - \alpha) S_2) \\ &\geq \min_{T \in U} [\alpha \omega(E, S_1) + (1 - \alpha) \omega(E, S_2)] \\ &\geq \alpha \min_{T \in U} \omega(E, S_1) + (1 - \alpha) \min_{T \in U} \omega(E, S_2) \end{aligned}$$

which means

$$\phi(\alpha S_1 + (1 - \alpha) S_2) \geq \alpha \phi(S_1) + (1 - \alpha) \phi(S_2)$$

Also by Lemma 1,

$$\lambda_{\min}(\alpha S_1 + (1 - \alpha) S_2) \geq \alpha \lambda_{\min}(S_1) + (1 - \alpha) \lambda_{\min}(S_2)$$

Combining the two inequalities yields

$$\begin{aligned} \psi_2(\alpha S_1 + (1 - \alpha) S_2) \\ &= \min\{\phi(\alpha S_1 + (1 - \alpha) S_2), \xi \lambda_{\min}(\alpha S_1 + (1 - \alpha) S_2)\} \\ &\geq \min\{\alpha \phi(S_1) + (1 - \alpha) \phi(S_2), \\ &\quad \alpha \xi \lambda_{\min}(S_1) + (1 - \alpha) \xi \lambda_{\min}(S_2)\} \\ &\geq \alpha \min\{\phi(S_1), \xi \lambda_{\min}(S_1)\} \\ &\quad + (1 - \alpha) \min\{\phi(S_2), \xi \lambda_{\min}(S_2)\} \\ &= \alpha \psi_2(S_1) + (1 - \alpha) \psi_2(S_2) \\ &> \psi_2(S_1), \end{aligned}$$

which contradicts the assumption that S_1 is a local maximum since α can be made arbitrarily close to 1. The proof is therefore completed

Theorem 3 in principle has allowed one to compute the maximum of $\psi_2(S)$ by a standard minimization algorithm

Theorem 1 can be rephrased in terms of the function $\psi_2(\cdot)$ as follows

Corollary 1 System (1) satisfying Assumption 1 is quadratically stabilizable if and only if

$$\eta_2 = \max_S \psi_2(S) > 0, \quad (12)$$

where ξ is an arbitrary positive scalar

Proof. The maximum is always defined, since W is clearly compact and $\psi_2(\cdot)$ is continuous. If (12) holds, let S^* be the maximum point, then (5) clearly holds for $S = S^*$ [notice that (12) implies $\lambda_{\min}(S^*) > 0$ and therefore $S^* \in S_1^*$], therefore the condition is sufficient. On the other hand, if the system is quadratically stabilizable, then (5) holds for some $S = \tilde{S}$. The matrix \tilde{S} can be rescaled to satisfy

$$\text{tr}(\tilde{S}) = 1$$

For such an \tilde{S} ,

$$\phi(\tilde{S}) > 0,$$

and

$$\lambda_{\min}(\tilde{S}) > 0$$

Therefore (12) holds. The condition is also necessary

The parameter ξ is meant to reflect the fact that S and $T_{A+T}^P S$ are generally of different scale. The design procedure for a stabilizing controller is as follows

1. Choose \tilde{D} , compute B and \tilde{D} as in Theorem 1. Choose ξ to reflect proper scaling due to the multiplication of \tilde{D} and A .
2. Maximize the function $\psi_2(S)$. Record the maximizing point S and the maximum η_2 . If $\eta_2 > 0$, the system is declared not quadratically stabilizable. Otherwise, proceed with step 3.
3. Choose a positive definite R . Let $P = S^{-1}$. Choose a sufficiently small ϵ such that (9) is satisfied. The stabilizing controller is

$$K = B^{-1} P$$

Let $\sigma_{\max}(Q)$ represent the maximum singular value of the matrix Q . Let

$$\begin{aligned} \sigma_R &= \sigma_{\max}(R), \\ \sigma_{\tilde{D}} &= \sigma_{\max}(\tilde{D}), \\ \sigma_1 &= \max_{T \in U} \sigma_{\max}(T_{A+T}^P S), \\ \sigma_R &= \sigma_{\max}(R) \end{aligned}$$

A choice of ϵ such that (9) is satisfied is

$$\epsilon = \frac{1}{\sigma_R(\sigma_R \sigma_1 + (\sigma_R \sigma_1 \sigma_{\tilde{D}})^2 / \eta_2)}$$

The set of uncertainty Ω is crucial in evaluating the function ψ_2 , which involves a minimization process. As mentioned above, when Ω is a hyperpolyhedron, [this is often the case in practice, for example, the uncertain system can often be expressed in the form of (2) with each uncertain factor bounded by $q_i \leq q_i(t) \leq \bar{q}_i$], the minimum is always reached by one of its vertices according to Theorem 2. It is therefore sufficient to compute the values at the vertices and take the minimum value. For other cases, we suggest to bound Ω by a hyperpolyhedron.

The maximization of $\psi_2(S)$ is computationally intensive. Fortunately, Theorem 3 points out that any local maximum is also a global maximum. Here a simple version of coordinate rotating algorithm (Rao, 1984) is described. This process is similar to that used in Gu *et al.* (1989). Express

$$S = U^T \Lambda U$$

where $U = (u_1, \dots, u_n)$ is an orthogonal matrix, and

$$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$$

Then the increment of coordinates can be realized by

changing

$$\lambda_i \rightarrow \lambda_i + \delta, \quad \lambda_n \rightarrow \lambda_n - \delta \quad \text{subject to } \lambda_i \geq 0, \quad \lambda_n \leq 0 \quad (13)$$

for some $i \in \{1, 2, \dots, n-1\}$ or

$$(u_i, u_j) \rightarrow (u_i \cos \theta - u_j \sin \theta, u_i \sin \theta + u_j \cos \theta) \quad (14)$$

for some $i, j \in \{1, \dots, n\}$, $i \neq j$. The algorithm is summarized as follows

1. $i = 0$. Choose initial searching step δ_0 and θ_0 . Choose minimum step δ^* and θ^* . Choose initial $S = U^T \Lambda U$.

$$U = I; \quad \Lambda = \frac{1}{n} I$$

2. Alternately increment coordinates to try to increase $\psi(S)$ by using (13) and (14) for

$$\delta = +\delta_i, \quad \theta = +\theta_i$$

3. If $\psi(S)$ has been increased in one round of step 2, repeat step 2. Otherwise, if $\delta_i = \delta^*$ and $\theta_i = \theta^*$, go to step 4. Otherwise, let $\delta_{i+1} = \delta_i/2$, $\theta_{i+1} = \theta_i/2$, $i = i + 1$; go to step 2.

4. Maximum η_i and maximum point S^* have been found. stop.

Other more advanced constrained optimization algorithm can also be used, see for example Rao (1984) and Polak (1987).

4. Generalizations

When both system matrix A and input matrix B are subject to uncertainties, a result due to Barmish (1983) can be applied:

Theorem 4. System (1) is quadratically stabilizable by a linear feedback control if and only if the following *augmented system* is quadratically stabilizable by a linear feedback control:

$$\frac{d}{dt} \begin{pmatrix} x \\ u \end{pmatrix} = \begin{pmatrix} A + \Delta A(t) & B + \Delta B(t) \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} + \begin{pmatrix} 0 \\ I \end{pmatrix} v$$

For this augmented system, the function $\phi(S)$ becomes

$$\phi(S) = \min_{(A, E) \in \Omega} [\lambda_{\min} \{ (A + E)S_{11} + (B + F)S_{12} - S_{11}(A + E)^T - S_{12}^T(B + F)^T \}]$$

where $S \in S^{n+n}$, and S_0 's are submatrices of S of commensurate dimensions. After the S maximizing $\psi_2(S)$ has been obtained, the stabilizing control is

$$u = -S_{12}S_{11}^{-1}x$$

The corresponding P matrix in Definition 1 is S_{11}^{-1} .

Another generalization is when Ω is not convex. For this case, since $\Omega \subset \text{conv}(\Omega)$, it is easily seen that the quadratic stabilizability by a linear control is implied by that of the same system with the uncertain set being replaced by $\text{conv}(\Omega)$. However, from Theorems 2 and 4, this is equivalent to the quadratic stabilizability by a linear control with uncertain set being $\text{prot}(\text{conv}(\Omega))$. Since

$$\text{prot}(\text{conv}(\Omega)) \subset \Omega,$$

we can reach the conclusion that the quadratic stabilizability by a linear control of a system (1) with uncertain set Ω is equivalent to that with uncertain set $\text{conv}(\Omega)$.

5. Illustrative example

A quadratic stabilizability by a linear feedback of the following system [a special case of this system was discussed by Petersen (1985)] is considered:

$$\dot{x} = \begin{pmatrix} r_1(t) & 0 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 1 \\ -r_2(t) \end{pmatrix} u,$$

$$(r_1(t), r_2(t)) \in \Omega \equiv \{[r_{1 \min}, r_{1 \max}] \times [r_{2 \min}, r_{2 \max}]\}$$

The system is quadratically stabilizable by a linear control if

and only if

$$\begin{pmatrix} r_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} + \begin{pmatrix} 1 \\ -r_2 \end{pmatrix} (s_{11} - s_{22}) + \text{transpose} \in S^2 \quad \text{for all } (r_1, r_2) \in \Omega$$

for some $S \in S^4$. After some algebra, it can be shown that such an S exists if and only if

$$4\sqrt{r_{1 \min}}\sqrt{r_{2 \min}} < (\sqrt{r_{1 \max}} - \sqrt{r_{1 \min}})(\sqrt{r_{2 \max}} - \sqrt{r_{2 \min}}).$$

The elements of S can be chosen to be

$$s_{11} = -1,$$

$$s_{22} = (\sqrt{r_{2 \max}} + \sqrt{r_{2 \min}})^2/4,$$

$$s_{12} = \frac{\sqrt{r_{2 \max}} + \sqrt{r_{2 \min}}}{\sqrt{r_{1 \max}} + \sqrt{r_{1 \min}}},$$

$$s_{11} = \frac{1}{2} \min \left(1 - \frac{(s_{21} + r_2 + s_{12}r_1)^2}{4s_{21}r_1r_2} \right),$$

$$s_{22} = (s_{12}^2 + 1)/s_{11}$$

and the stabilizing control is

$$u = (s_{11} - s_{22}) \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}$$

6. Conclusions

The necessary and sufficient conditions for quadratic stabilizability of uncertain linear systems by linear feedback control are formulated in a form of two level optimization problem. It is found that when the uncertainty region is a hyperpolyhedron, the optimization of inner level is reduced to comparing the function values at the corners. It is proposed that arbitrary uncertain regions be bounded by a hyperpolyhedron and therefore the proposed algorithm can be applied. It is also proved that any local maximum in the outer level optimization is also the global maximum, and therefore any standard optimization algorithms can be applied.

References

- Barmish, B. R. (1983). Stabilization of uncertain systems via linear control. *IEEE Trans. Aut. Control*, **28**, 848-850.
- Barmish, B. R. (1985). Necessary and sufficient conditions for quadratic stabilizability of an uncertain system. *J. Optimiz. Theory Applic.*, **46**, 399-408.
- Islami, M. and D. L. Russell (1980). On stability with large parameter variations. Stemming from the direct method of Lyapunov. *IEEE Trans. Aut. Control*, **AC-25**, 1231-1234.
- Gu, K., M. A. Zohdy and N. K. Loh (1989). Necessary and sufficient conditions of quadratic stability of uncertain linear systems, *28th IEEE Conf. on Decision and Control*, Tampa, FL; also *IEEE Trans. Aut. Control*, 1990, **AC-35**, 601-604.
- Hollot, C. V. and B. R. Barmish (1980). Optimal quadratic stabilizability of uncertain linear systems. *Proc. of the 18th Allerton Conf. on Communication, Control and Computing*, University of Illinois, Monticello, IL.
- Khargonekar, P. P., I. R. Petersen and K. Zhou (1988). Robust stabilization of uncertain linear systems: Quadratic stability and H^∞ control theory. *IEEE Trans. Aut. Control*, **AC-35**, 356-361.
- Luenberger, D. (1969). *Optimization by Vector Space Methods*, Wiley, New York.
- Patel, R. V., M. Toda and B. Sridhar (1977). Robustness of linear quadratic state feedback designs in the presence of the system uncertainty. *IEEE Trans. Aut. Control*, **AC-22**, 945-949.
- Patel, R. V. and M. Toda (1980). Quantitative measures of robustness for multivariable systems. *Proc. Joint Aut. Control Conf.*, San Francisco, CA, paper TD8-A.
- Petersen, I. R. (1985). Quadratic stabilizability of uncertain linear systems. Existence of a nonlinear stabilizing control

- does not imply existence of a linear stabilizing control. *IEEE Trans. Aut. Control*, **30**, 291-293.
- Petersen, I. R. (1987). A stabilization algorithm for a class of uncertain linear systems. *Syst. Control Lett.*, **8**, 351-357.
- Petersen, I. R. (1988). Stabilization of an uncertain linear system in which uncertain parameters enter into the input matrix. *SIAM J. Control Optimiz.*, **26**, 1257-1263.
- Polak, E. (1987). On the mathematical foundations of nondifferentiable optimization in engineering design. *SIAM Rev.*
- Rao, S. S. (1984). *Optimization Theory and Applications*. Wiley, New York.
- Yedavalli, R. K. (1985a). Improved measure of stability robustness for linear state space models. *IEEE Trans. on Aut. Control*, **AC-30**, 577-579.
- Yedavalli, R. K. (1985b). Perturbation bounds for robust stability in linear state space models. *Int. J. Control*, **42**, 1507-1517.
- Yedavalli, R. K. (1986). Stability robustness measures under dependent uncertainty. 1986 *Amer. Control Conf.*, Atlanta, GA, pp. 820-823.
- Zhou, K. and P. P. Khargonekar (1987). Stability robustness for linear state-space models with structured uncertainty. *IEEE Trans. Aut. Control*, **AC-32**, 621-623.

Brief Paper

Some Majorant Robustness Results for Discrete-time Systems*

DAVID C. HYLAND† and EMMANUEL G. COLLINS, Jr‡

Key Words—Robustness, sensitivity analysis, discrete-time systems

Abstract—This paper uses majorant bounding techniques to develop robust performance bounds and stability conditions for linear discrete-time systems expressed in a basis in which the state matrix is block-diagonal. Performance is measured in terms of the time-dependent behavior of the system outputs. This performance criterion is not considered explicitly by other robustness results. Two sets of robustness results are developed. The second set is able to capture the intuitive effect that increasing the distance between the subsystems increases the magnitude of nondestabilizing uncertain coupling allowed between the subsystems. In addition, the developments demonstrate that alternative representations of a given uncertain system can lead to significantly different robustness bounds.

1 Introduction

MAJORANT BOUNDING techniques were developed by Dahlquist to produce bounds for the solutions of systems of equations (Dahlquist, 1983). Similar bounding techniques have been used in the work of researchers in large-scale systems analysis (Porter and Michel, 1974; Lasley and Michel, 1976). More recently, majorant bounding techniques have been used to develop robust stability and performance results for the analysis of linear time-invariant systems with structured uncertainty (Hyland and Bernstein, 1987; Collins and Hyland, 1989; Hyland and Collins, 1989a). Previous results in majorant robustness analysis allow system performance to be measured in two ways. The initial developments (Hyland and Bernstein, 1987; Collins and Hyland, 1989) allow the performance to be measured in terms of the deviations of the steady-state variances of selected system variables from their nominal values. Later developments in the frequency domain (Hyland and Collins, 1989a) allow the performance to be measured in terms of frequency-dependent deviations of the system outputs from their nominal values. The present paper also uses majorant bounding techniques to develop robust analysis results. Specifically, it considers the case in which the performance is measured in terms of the time-dependent deviations of selected system outputs from their nominal values. This allows analysis of both the transient and steady state behaviors of the system. The systems analyzed are a class of discrete-time linear time-invariant systems which have structured uncertainty and are expressed in a basis in which the state matrix is block-diagonal. The systems are subjected to deterministic signals. This is compatible with standard classical analysis which often injects specified signals (i.e. steps, ramps and sinusoids) to measure system behavior.

Two sets of results are developed. The second set is able to capture the intuitive effect that increasing the distance between the spectra of two subsystems increases the magnitude of nondestabilizing uncertain coupling allowed between the subsystems. This phenomena is not observable by many standard analysis tools such as vector and quadratic Lyapunov theories and singular value analysis. In addition, the developments here demonstrate that alternative representations of a given uncertain system can lead to significantly different robustness bounds.

The paper is organized as follows. Section 2 gives notation and some important mathematical relations. Section 3 presents and formulates a problem in robust stability and performance analysis for discrete-time systems with block-structured uncertainty. Section 4 then presents preliminary developments. The main results are found in Section 5 which uses majorant bounding techniques to develop robust performance bounds and sufficient conditions for robust stability. Section 6 illustrates the results with a numerical example. Finally, Section 7 presents concluding remarks.

2 Notation and mathematical relations

\mathbb{N}, \mathbb{N}_0	set of (nonnegative) integers
\mathbb{R}, \mathbb{R}_0	set of (nonnegative) real numbers
I_p	$p \times p$ identity matrix
\odot	$\{y_{ij}, z_{ij}\}$, Hadamard product of matrices
$Y \cdot Z$	Y, Z of equal dimensions (Styan, 1973)
\otimes, \oplus	Kronecker product and sum (Brewer, 1978)
$\text{vec}(Z)$	column stacking operator associated with the Kronecker product (Brewer, 1978)
$Y \sim Z$	$y_{ij} = z_{ij}$ for each i and j (Y and Z are real matrices of equal dimension)
$ Z _u$	modulus matrix of $Z (= z_{ij})$
$\lambda(Z), \rho(Z)$	eigenvalue and spectral radius of square matrix Z [$\ Z\ > 0, \rho(Z)$ is the Perron root of Z (Seneta, 1973; Herman and Plenum, 1979)]
$\ \cdot \ _2, \ \cdot \ _F, \ \cdot \ _1, \ \cdot \ _\infty$	Euclidean, spectral and Frobenius norms

Let $Y, Z \in \mathbb{C}^{n \times n}$ and $u, v \in \mathbb{C}^n$ have the partitioned forms

$$Y = [Y_{ij}]_{i,j=1}^4, \quad Z = [Z_{ij}]_{i,j=1}^4, \quad u = [u_i^T]_{i=1}^4 \quad (2.1, 2, 3, 4)$$

where $Y_{ij}, Z_{ij} \in \mathbb{C}^{n_i \times n_j}$, $u_i \in \mathbb{C}^{n_i}$ and $\sum_{i=1}^4 n_i = n$. The block-norm matrix of Z with respect to the matrix norm $\| \cdot \|_p$ (Oktrowski, 1961) (which is also called the " θ block norm") is the $r \times r$ nonnegative matrix

$$Z_\theta = [Z]_\theta \triangleq [\|Z_{ij}\|_p]_{i,j=1}^4 \quad (2.5)$$

Thus, Z_s and Z_F represent respectively the block-norm matrices of Z with respect to the spectral and Frobenius norms. The block-norm vector of u with respect to the Euclidean norm (or "Euclidean block-norm") is the nonnegative vector $u \in \mathbb{R}^r$ defined by

$$u^T \triangleq [\|u_1\|_2, \dots, \|u_4\|_2] \quad (2.6)$$

Notice that if the partitions of the matrix Z and vector u are their scalar elements (i.e. $r = n$), then the spectral (or

* Received 24 October 1988, revised 13 July 1989, received in final form 7 February 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak.

† Harris Corporation, Government Aerospace Systems Division, MS 22/4842, Melbourne, FL 32902, U.S.A.

‡ Author to whom all correspondence should be addressed.

Frobenius) block-norm matrix and block-norm vector are simply modulus matrices. That is, $Z_N = Z_F = \|Z\|_M$ and $u = \|u\|_M$ for $r = n$.

Subsequent analysis will also use the following block-norm relations (Hyland and Collins, 1989a)

$$\begin{aligned} \|(u \otimes v) \otimes z, u \otimes v, (Zu) \otimes z\|_M &\leq \|Z\|_M \|u\|_M \|v\|_M \\ \|(YZ) \otimes z\|_M &\leq \|Y\|_M \|Z\|_M \|z\|_M \\ \|Z\|_M &\leq \|Z\|_F, \|u \otimes v\|_M \leq \|u\|_F \|v\|_F \end{aligned} \tag{2.7, 8, 9}$$
$$\|Z\|_M \leq \|Z\|_F, \|u \otimes v\|_M \leq \|u\|_F \|v\|_F \tag{2.10, 11}$$

The analysis will also require the matrix norm relations

$$\begin{aligned} \|A\|_M &\leq \|A\|_F, \|AB\|_F \leq \|A\|_M \|B\|_F \tag{2.12, 13} \\ \|A\|_F &= \|\text{vec } A\|_F, \|A \otimes B\|_M \leq \|A\|_M \|B\|_M \tag{2.14, 15} \end{aligned}$$

where A and B are arbitrary complex matrices of compatible dimension. [A proof of (2.15) is presented in Hyland and Collins (1989b)]

Majorants (Dahlquist, 1983) are essentially element-by-element upper bounds for block-norms. Precisely, $Z_n \in \mathbb{R}_+^{n \times n}$ is a *majorant* of Z with respect to the 0 matrix norm (or “0 majorant”) if $Z_n \geq \|Z\|_0$. Similarly, $\tilde{u} \in \mathbb{R}_+^n$ is a *majorant* of u with respect to the Euclidean norm (or “Euclidean majorant”) if $\tilde{u} \geq \|u\|_2$.

A matrix $P \in \mathbb{R}^{n \times n}$ is an *M-matrix* (Fiedler and Ptak, 1962; Seneta, 1973; Berman and Plemmons, 1979) if it has nonpositive off-diagonal elements (i.e. $p_{ij} \leq 0$ for $i \neq j$) and positive principal minors. It is shown in Berman and Plemmons (1979) that if $D \in \mathbb{R}_+^{n \times n}$, then $L_n - D$ is an *M-matrix* if and only if $\rho(D) < 1$.

Finally, note that if $Z \in \mathbb{C}^{n \times n}$ and $(I - Z)$ is nonsingular, then

$$\sum_{i=0}^{\infty} Z^i = (I - Z^{-1})(I - Z) \tag{2.16}$$

This relation is easily proved by postmultiplying both sides of (2.16) by $(I - Z)$ and expanding the left hand side

3. Problem formulation

Consider the discrete-time system

$$x(k+1) = (A+G)x(k) + v(k), \quad x(0) \in \mathbf{X} \tag{3.1}$$

where $x \in \mathbb{C}^n$, $\mathbf{X} \subset \mathbb{C}^n$ and v is a known input which is assumed to be bounded for all time (i.e. $\sup_{k \geq 0} \|v(k)\| < \infty$). It is assumed that the system (3.1) is comprised of r ($r \leq n$) nominally stable subsystems described by A with uncertain interactions and dynamics described by G . Specifically,

$$A = \text{block diag } \{A_i\}_i \tag{3.2}$$

where $A_i \in \mathbb{C}^{n_i \times n_i}$ is asymptotically stable (i.e. all eigenvalues lie in the open unit circle) and $\sum_{i=1}^r n_i = n$. The subsystems described by the A_i can be either physical subsystems (e.g. spatially distinct parts of a complex flexible structure) or mathematical subsystems (e.g. the modes of the nominal system). The matrix $G \in \mathbb{C}^{n \times n}$ and the vectors $x, v \in \mathbb{C}^n$ are partitioned conformably so that

$$G = [G_{ij}]_{i,j=1}^r, \quad [x_i^T, \quad v_i^T]^T \tag{3.3, 4, 5}$$

where $G_{ij} \in \mathbb{C}^{n_i \times n_j}$ and $x_i, v_i \in \mathbb{C}^{n_i}$.

In what follows it is assumed that for some $\hat{x}^* \in \mathbb{R}_+^n$ the initial condition set \mathbf{X} is described by

$$\mathbf{X} = \{x(0) \in \mathbb{R}_+^n; \hat{x}(0) \leq \hat{x}^*\} \tag{3.6}$$

We also consider two descriptions of the admissible uncertainty. Specifically, for some $\hat{G}_i \in \mathbb{R}_+^{n_i \times n_i}$ or $\hat{G}_i \in \mathbb{R}_+^{n_i \times n_i}$ it is assumed that either $G \in \mathbf{G}_s$ or $G \in \mathbf{G}_f$ where

$$\begin{aligned} \mathbf{G}_s &= \{G \in \mathbb{C}^{n \times n}; \hat{G}_i \leq \|G_i\|_M\} \\ \mathbf{G}_f &= \{G \in \mathbb{C}^{n \times n}; \hat{G}_i \leq \|G_i\|_F\} \end{aligned} \tag{3.7, 8}$$

That is, the admissible uncertainty is described either in terms of upper bounds on the *spectral norms* of the subblocks of G or in terms of upper bounds on the *Frobenius norms* of the subblocks of G .

The performance of the system (3.1) is measured in terms of the time-dependent behavior of the scalar output $y(\cdot)$ given by

$$y(k) = c^T x(k) \tag{3.9}$$

The partitioned form of $c \in \mathbb{R}^n$ is given by

$$c^T = [c_1^T, \dots, c_r^T] \tag{3.10}$$

where $c_i \in \mathbb{R}^{n_i}$.

Now assume \mathbf{G} represents the uncertainty set \mathbf{G}_s or \mathbf{G}_f and define

$$y_{\min}(k) \triangleq \min_{x(0) \in \mathbf{X}, G \in \mathbf{G}} y(k), \quad y_{\max}(k) \triangleq \max_{x(0) \in \mathbf{X}, G \in \mathbf{G}} y(k), \quad k \geq 0 \tag{3.11, 12}$$

Remark 1. It should be noted that both $y_{\min}(k)$ and $y_{\max}(k)$ are implicit functions of the system input $v(\cdot)$. Thus, a more precise notation would be $y_{\min}[v(\cdot), k]$ and $y_{\max}[v(\cdot), k]$. However for the sake of simplicity this notation is suppressed.

Definition 1. The functions $\alpha: \mathbb{N}_0 \rightarrow \mathbb{R}$ and $\beta: \mathbb{N}_0 \rightarrow \mathbb{R}$ are respectively *lower* and *upper performance bounds* with respect to the uncertainty set \mathbf{G} if $y_{\min}(k) \geq \alpha(k)$ and $y_{\max}(k) \leq \beta(k)$.

Definition 2. If $A + G$ is asymptotically stable for all $G \in \mathbf{G}$, then the system (3.1) is *robustly stable* over \mathbf{G} .

The aim of subsequent developments is twofold. First, we will develop lower and upper performance bounds for the system (3.1). In addition, we will determine sufficient conditions for the robust stability of (3.1). It should be noted that although the performance analysis given below explicitly considers only the scalar output (3.9), the results actually allow *multiple objective analysis*. That is other (scalar) outputs which are linear functions of the state can be analyzed by simply changing the vector c in (3.9).

4. Preliminary developments

Before presenting the main results in Section 5 some additional development is required.

Consider the nominal system

$$x^0(k+1) = Ax^0(k) + v(k), \quad x^0(0) = x(0) \tag{4.1}$$

and note that the state vector $x^0(\cdot)$ can be expressed as

$$x^0(k) = \sum_{i=0}^{k-1} A^i v(i) + A^k x(0), \quad x^0(0) = z(0) \tag{4.2}$$

Also, define the sets

$$\begin{aligned} \mathbf{X}^0 &\triangleq \{x^0(\cdot); x(0) \in \mathbf{X}\} \\ \mathbf{X}_i^0 &\triangleq \{x^0(\cdot); x^0(k) = A^k x(0), x(0) \in \mathbf{X}\} \end{aligned} \tag{4.3, 4}$$

\mathbf{X}^0 is the *set of nominal responses* while \mathbf{X}_i^0 is the *set of nominal initial condition responses*. Obviously, if $v(k) = 0$ for $k \geq 0$, then $\mathbf{X}^0 = \mathbf{X}_i^0$.

Next, define

$$e(k) \triangleq x(k) - x^0(k) \tag{4.5}$$

Then, subtracting (4.1) from (3.1) gives

$$e(k+1) = (A+G)e(k) + Gx^0(k), \quad e(0) = 0 \tag{4.6}$$

The partitioned form of e is

$$e^T = [e_1^T, \dots, e_r^T] \tag{4.7}$$

where $e_i \in \mathbb{C}^{n_i}$.

Definition 3. The function $\hat{e}: \mathbb{N}_0 \rightarrow \mathbb{R}_+^r$ is a *perturbation bound* with respect to the reference input $v(\cdot)$ if for each $x^0(\cdot) \in \mathbf{X}$ and $G \in \mathbf{G}$ the resultant error sequence $\{e(k)\}_{k=0}^{\infty}$ satisfies

$$e(k) \leq \hat{e}(k) \tag{4.8}$$

Now, the output $y(\cdot)$ can be expressed as

$$y(k) = y^0(k) + \Delta y(k) \quad (4.9)$$

where

$$y^0(k) = c^T x^0(k), \quad \Delta y(k) = c^T e(k) \quad (4.10, 11)$$

The following lemma uses the above characterization of $y(\cdot)$ to present lower and upper performance bounds for the system (3.1).

Lemma 1. Consider the system (3.1) and suppose $\hat{e}(\cdot)$ is a perturbation bound with respect to the reference input $v(\cdot)$ and let

$$\bar{\alpha}(k) = y^0(k) - \hat{c}^T \hat{e}(k), \quad \bar{\beta}(k) = y^0(k) + \hat{c}^T \hat{e}(k) \quad (4.12, 13)$$

Then, $\bar{\alpha}: \mathbb{N}_0 \rightarrow \mathbb{R}$ and $\bar{\beta}: \mathbb{N}_0 \rightarrow \mathbb{R}$ are respectively lower and upper performance bounds with respect to the reference input $v(\cdot)$.

Proof. From (4.11), (2.11) and (4.8) it follows that $|\Delta y(k)| \leq \hat{c}^T \hat{e}(k)$ for each $G \in \mathbf{G}$. The proof is then immediate from (4.9).

Robust stability can be determined by considering the homogeneous system

$$x(k+1) = (A+G)x(k), \quad x(0) = x(0) \quad (4.14)$$

The following proposition gives an equivalent condition for robust stability.

Proposition 1. The system (3.1) is robustly stable over \mathbf{G} if and only if for any strictly positive initial condition block-norm bound \bar{x}^* , characterizing the set \mathbf{X} defined by (3.6), the state $x(\cdot)$ of the homogeneous system (4.14) satisfies $\lim_{k \rightarrow \infty} \|x(k)\| = 0$ for each $x(0) \in \mathbf{X}$ and $G \in \mathbf{G}$.

Now, assume that $v(k) = 0$ for $k \geq 0$ and thus $x^0(\cdot) \in \mathbf{X}_f^*$. In this case, $\lim_{k \rightarrow \infty} \|x^0(k)\| = 0$ and $e(k) = x(k) - x^0(k)$. It follows that the robust stability condition of Proposition 1 is equivalent to $\lim_{k \rightarrow \infty} \|e(k)\| = 0$ for $x(0) \in \mathbf{X}$ and $G \in \mathbf{G}$. The next lemma used in the following section to develop sufficient conditions for robust stability, is an immediate consequence of the above equivalence and Proposition 1.

Lemma 2. Assume that $\bar{x}^* > 0$, $v(k) \geq 0$ (i.e. $x^0(\cdot) \in \mathbf{X}_f^*$) and $\hat{e}(\cdot)$ is a perturbation bound for the system (3.1). Then, if $\lim_{k \rightarrow \infty} \hat{e}(k) = 0$, the system (3.1) is robustly stable.

From the above discussion it is apparent that an important part of the ensuing results is the development of perturbation bounds for the system (3.1). With this in mind, we present two expressions for the solution $e(\cdot)$ of (4.6) which will later be used to derive perturbation bounds.

It is well known that the solution $e(\cdot)$ of (4.6) can be expressed as

$$e(k) = \sum_{l=0}^{k-1} A^{k-l-1} G [e(l) + x^0(l)], \quad e(0) = 0 \quad (4.15)$$

This is the first expression for $e(\cdot)$. Substituting (4.15) into its own right hand side and utilizing the substitution

$$\sum_{l=0}^{k-1} \sum_{m=0}^{l-1} = \sum_{m=0}^{k-2} \sum_{l=m+1}^k \quad (4.16)$$

then gives

$$e(k) = \sum_{m=0}^{k-2} \mathcal{M}(k-1, m+1) G [e(m) + x^0(m)] + \sum_{l=0}^{k-1} A^{k-l-1} G x^0(l), \quad e(0) = 0 \quad (4.17)$$

where

$$\mathcal{M}(k-1, m+1) \triangleq \sum_{l=m+1}^k A^{k-l-1} G A^{l-m-1} \quad (4.18)$$

Equation (4.17) is the second expression for $e(\cdot)$. If G were known and one were actually numerically solving for $e(\cdot)$, (4.15) and (4.17) would yield the same solution. However, since G is unknown, (4.15) and (4.17) will be used to

produce perturbation bounds for the system (3.1). As is seen later, bounds derived from the two equations can differ significantly.

5. Robust stability and performance analysis

This section presents robust performance bounds and sufficient conditions for robust stability of the system (3.1). The main results are found in Theorems 1 and 2. These theorems are dependent respectively on Lemmas 4 and 5 which give perturbation bounds for the system (3.1). Theorem 4 compares the previous results.

We begin the presentation by stating an important comparison lemma.

The proof of this lemma is trivial and is thus omitted.

Lemma 3. For each $l \geq 0$ let $y^0(l) \in \mathbb{R}^{n_y}$ and for each $k \geq 0$ let $u^{*k} \in \mathbb{R}^m$. Also, let $f: \mathbb{N}_0 \rightarrow \mathbb{R}^n$ and $g: \mathbb{N}_0 \rightarrow \mathbb{R}^n$ satisfy

$$f(k) = \sum_{l=0}^k y^0(l) + u^{*k}, \quad f(0) = f^* \quad (5.1)$$

$$g(k) = \sum_{l=0}^k y^0(l) + u^{*k}, \quad g(0) = g^* \quad (5.2)$$

where $f^*, g^* \in \mathbb{R}^n$. Then, if $f^* \leq g^*$,

$$f(k) \leq g(k) \quad (5.3)$$

Before presenting the next result, we present a remark to help clarify notation.

Remark 2. Below we use the notation

$$A_k^{-1} \triangleq (A^k)^{-1} \quad (5.4)$$

Proposition 2. Let $x^0(\cdot)$ be given by

$$x^0(k+1) = A_k x^0(k) + v(k), \quad x(0) = x^* \quad (5.5)$$

Then

$$x^0(k) = -x^0(k), \quad x(0) \in \mathbf{X} \quad (5.6)$$

Proof. Since $x(0) \in \mathbf{X}$, $x(0) \leq x^*$. By using (2.9), (2.11) and (3.6) it then follows from (4.2) that for all $x(0) \in \mathbf{X}$

$$x^0(k) = x^0(k) + \sum_{l=0}^{k-1} A_k^{k-l-1} v(l) + A_k^k x^* \quad (5.7)$$

The proof is completed by simply recognizing the equivalence of (5.5) and the equality in (5.7). \square

In Lemma 4 and Theorem 1 presented below, it is assumed that \mathbf{G} and \mathbf{G}_γ represent either \mathbf{G}_γ and \mathbf{G}_γ or \mathbf{G}_γ and \mathbf{G}_γ .

Lemma 4. Let $\hat{e}^{(1)}: \mathbb{N}_0 \rightarrow \mathbb{R}^n$ be given by

$$\hat{e}^{(1)}(k+1) = (A_k + G) \hat{e}^{(1)}(k) + G x^0(k), \quad \hat{e}^{(1)}(0) = 0 \quad (5.8)$$

Then, $\hat{e}^{(1)}(\cdot)$ is a perturbation bound of the system (3.1) with respect to the uncertainty set \mathbf{G} .

Proof. First assume that $G \in \mathbf{G}_{\gamma_1}$ such that $G_{\gamma_1} \leq G_{\gamma_2}$. By using (2.9), (2.11) and (5.6) it then follows from (4.15) that for each $x(0) \in \mathbf{X}$ and $G \in \mathbf{G}_{\gamma_1}$

$$e(k) = \sum_{l=0}^{k-1} A_k^{k-l-1} G_{\gamma_1} [e(l) + x^0(l)], \quad e(0) = 0 \quad (5.9)$$

Now recognize that (5.8) is equivalent to

$$\hat{e}^{(1)}(k) = \sum_{l=0}^{k-1} A_k^{k-l-1} G_{\gamma_1} [\hat{e}^{(1)}(l) + x^0(l)], \quad \hat{e}^{(1)}(0) = 0 \quad (5.10)$$

From (5.9), (5.10) and Lemma 3 it follows that for all $x(0) \in \mathbf{X}$ and $G \in \mathbf{G}_{\gamma_1}$, $e(k) \leq \hat{e}^{(1)}(k)$.

Next assume $G \in \mathbf{G}_{\gamma_2}$ such that $G_{\gamma_2} \leq G_{\gamma_1}$. It then follows from (2.10) that $G_{\gamma_2} \leq G_{\gamma_1}$. The proof is concluded by repeating the above steps using $G_{\gamma_2} \leq G_{\gamma_1}$ instead of $G_{\gamma_1} \leq G_{\gamma_2}$. \square

Theorem 1. Consider the system (3.1) with reference input $v(\cdot)$ and let

$$\alpha^{(1)}(k) = y^0(k) - \hat{c}^T \hat{e}^{(1)}(k), \quad \beta^{(1)}(k) = y^0(k) + \hat{c}^T \hat{e}^{(1)}(k) \quad (5.11, 12)$$

where $\hat{e}^{(1)}(\cdot)$ is defined by (5.8). Then, $\alpha^{(1)}(\cdot)$ and $\beta^{(1)}(\cdot)$ are respectively lower and upper performance bounds with respect to the uncertainty set \mathbf{G} . In addition, if

$$\rho(\tilde{A}_s + \tilde{G}) < 1, \quad (5.13)$$

the system (3.1) is robustly stable.

Proof. That $\alpha^{(1)}(\cdot)$ and $\beta^{(1)}(\cdot)$ are performance bounds follows immediately from Lemma 4 and Lemma 1. Now assume that (5.13) holds. Then, if the bound \hat{x}^* for the block-norm of the initial state is strictly positive (i.e. $\hat{x}^* \gg 0$) and $v(k) \hat{=} 0$ (i.e. $x^0(\cdot) \in \mathbf{X}_v^0$), it follows from (5.8) that $\lim_{k \rightarrow \infty} \hat{e}^{(1)}(k) = 0$. Thus, using Lemma 2 it follows that (5.13) is a sufficient condition for robust stability of the system (3.1). \square

Remark 3. The robust stability condition (5.13) can be very conservative since even if $\tilde{G} = 0$ it is possible that for stable A_i , $\rho(A_i) > 1$. However, if the subsystems of A correspond to the modes of the system this phenomena does not exist. In general, the smaller the subsystems are the less conservative are the robustness results of Lemma 4 and Theorem 1.

The robustness results presented above were developed by using the characterization of $e(\cdot)$ given by (4.15) and considered the uncertainty sets \mathbf{G}_s and \mathbf{G}_p . The next results depend on the characterization of $e(\cdot)$ given by (4.17) and consider only the uncertainty set \mathbf{G}_p .

Before presenting these results for $k > 0$ and $l > 0$ define the matrix $\mathcal{P}(k-l-1)$ by

$$\mathcal{P}(k-l-1) = [\mathcal{P}_{ij}(k-l-1)]_{(i,j)=1,\dots,r} \quad (5.14)$$

where $\mathcal{P}_{ij}(k-l-1)$ is the $n_i n_i \times n_j n_j$ matrix with block partitions,

$$\mathcal{P}_{ij}(k-l-1) = \sum_{s=0}^{k-l-2} (A_i^s)^T \otimes A_i^{k-l-2-s} \quad (5.15)$$

Here and below we use the convention that

$$\sum_{m=u}^v (\cdot) \hat{=} 0 \quad \text{for } v < u \quad (5.16)$$

Lemma 5. Let $\hat{e}^{(2)}: \mathbb{N}_0 \rightarrow \mathbb{R}^n$ be given by

$$\begin{aligned} \hat{e}^{(2)}(k) = & \sum_{l=0}^{k-2} [\mathcal{P}_s(k-l-1) * \tilde{G}_p] \tilde{G}_p [\hat{e}^{(2)}(l) + \hat{x}^0(l)] \\ & + \sum_{l=0}^{k-1} (\tilde{A}_s)^{k-l-1} \tilde{G}_p \hat{x}^0(l), \quad \hat{e}^{(2)}(0) = 0 \end{aligned} \quad (5.17)$$

Then, $\hat{e}^{(2)}(\cdot)$ is a perturbation bound of the system (3.1) with respect to the uncertainty set \mathbf{G}_p .

Proof. Using (2.7)–(2.9) with (5.6), it follows from (4.17) that

$$\begin{aligned} \hat{e}(k) \leq & \sum_{m=0}^{k-2} \mathcal{M}_s(k-1, m+1) \tilde{G}_s [\hat{e}(l) + \hat{x}^0(m)] \\ & + \sum_{m=0}^{k-1} (A_s)^{k-m-1} \tilde{G}_s \hat{x}^0(m), \quad \hat{e}(0) = 0 \end{aligned} \quad (5.18)$$

If $G \in \mathbf{G}_p$, then $G_p \leq \leq \tilde{G}_p$ which by using (2.10) implies $\tilde{G}_s \leq \leq \tilde{G}_p$. Now, using (2.12), (2.14), and (2.13) with some simple Kronecker product identities (Brewer, 1978) yields

$$\begin{aligned} \|\mathcal{M}_s(k-1, m+1)\|_s & \leq \left\| \sum_{i=1}^r (A_i)^T \otimes A_i^{k-m-1} \right\| \|\tilde{G}_s\|_p, \quad i, j = 1, \dots, r \\ & \quad (5.19) \end{aligned}$$

or equivalently

$$\mathcal{M}_s(k-1, m+1) \leq \leq \mathcal{P}_s(k-m-1) * \tilde{G}_p \quad (5.20)$$

where $\mathcal{P}_s(k-m-1)$ is given by (5.14) and (5.15). Substituting $\tilde{G}_s \leq \leq \tilde{G}_p$, (5.20), and $\tilde{G}_p \leq \leq \tilde{G}_p$ into (5.18)

gives

$$\begin{aligned} \hat{e}(k) \leq & \sum_{l=0}^{k-2} [\mathcal{P}_s(k-l-1) * \tilde{G}_p] \tilde{G}_p [\hat{e}(l) + \hat{x}^0(l)] \\ & + \sum_{l=0}^{k-1} (\tilde{A}_s)^{k-l-1} \tilde{G}_p \hat{x}^0(l), \quad \hat{e}(0) = 0. \end{aligned} \quad (5.21)$$

It then follows from (5.17), (5.21) and Lemma 3 that $\hat{e}(k) \leq \leq \hat{e}^{(2)}(k)$. \square

The next lemma shows that the calculation of $\mathcal{P}(\cdot)$ can often be considerably simplified which in turn simplifies the computation of the perturbation bound $\hat{e}^{(2)}(\cdot)$.

Lemma 6. Assume that for $i \in \{1, 2, \dots, r\}$ A_i is nonsingular and for $i \neq j$, A_i and A_j do not have common eigenvalues. Then, the partitions of $\mathcal{P}(\cdot)$ defined by (5.15) can be expressed as

$$\begin{aligned} \mathcal{P}_{ij}(k-l-1) & = \begin{cases} (I_{n_i} \otimes A_i^{k-l-2}) \sum_{p=0}^{l-1} (A_i^p \otimes A_i^{l-1-p})^T, & i=j \\ [(-A_i^T)^{k-l-1} \oplus A_i^{k-l-1}] [(-A_i^T) \oplus A_i]^{-1}, & i \neq j \end{cases} \end{aligned} \quad (5.22a, b)$$

Proof. See Appendix A. \square

Theorem 2. Consider the system (3.1) with reference input $v(\cdot)$ and the uncertainty set \mathbf{G}_p . Let

$$\alpha^{(2)}(k) = y^0(k) - \hat{e}^{(2)}(k), \quad \beta^{(2)}(k) = y^0(k) + \hat{e}^{(2)}(k) \quad (5.23, 24)$$

where $\hat{e}^{(2)}(\cdot)$ is defined by (5.17). Then, $\alpha^{(2)}(\cdot)$ and $\beta^{(2)}(\cdot)$ are respectively lower and upper performance bounds. In addition, if the initial condition block-norm bound \hat{x}^* is strictly positive (i.e. $\hat{x}^* \gg 0$), $v(k) \hat{=} 0$ (i.e. $x^0(\cdot) \in \mathbf{X}_v^0$) and $\lim_{k \rightarrow \infty} \hat{e}^{(2)}(k) = 0$, then the system is robustly stable.

Proof. That $\alpha^{(2)}(\cdot)$ and $\beta^{(2)}(\cdot)$ are performance bounds follows immediately from Lemma 4 and Lemma 1. The robust stability condition follows immediately from Lemma 5 and Lemma 2. \square

In the final theorem of this section it is seen that with respect to the uncertainty set \mathbf{G}_p Lemma 5 and Theorem 2 (at the expense of greater computational intensity) give less conservative robustness results than Lemma 4 and Theorem 1.

Theorem 3. Assume that $\mathbf{G} = \mathbf{G}_p$ in Lemma 4 and Theorem 1. Then, the perturbation bounds $\hat{e}^{(1)}(\cdot)$ and $\hat{e}^{(2)}(\cdot)$ defined respectively by (5.8) and (5.17) satisfy

$$\hat{e}^{(2)}(k) \leq \leq \hat{e}^{(1)}(k) \quad (5.25)$$

Thus, the corresponding lower performance bounds $\alpha^{(1)}(\cdot)$ and $\alpha^{(2)}(\cdot)$ defined by (5.11) and (5.23) and upper performance bounds $\beta^{(1)}(\cdot)$ and $\beta^{(2)}(\cdot)$ defined by (5.12) and (5.24) satisfy

$$\alpha^{(1)}(k) \leq \leq \alpha^{(2)}(k), \quad \beta^{(2)}(k) \leq \leq \beta^{(1)}(k). \quad (5.26, 27)$$

In addition, if the robust stability condition (5.13) of Theorem 1 is satisfied, then the robust stability condition given in Theorem 2 is also satisfied.

Proof. See Appendix B. \square

Remark 4. Suppose $r = 2$ and for $\gamma \in \mathbb{R}$,

$$\tilde{G}_p = \gamma \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad A_i = \begin{bmatrix} \mu_i & \theta_i \\ -\theta_i & \mu_i \end{bmatrix}, \quad i = 1, 2 \quad (5.28, 29)$$

such that

$$\lambda_{1,2}(A_i) = \mu_i \pm j\theta_i \quad (5.30)$$

Then, if $A_1 \neq 0$ or $A_2 \neq 0$, and either $\mu_1 \neq \mu_2$ or $\theta_1 \neq \theta_2$,

$$\| [(-A_i^T) \oplus A_i]^{-1} \|_s = \left[\frac{1}{(\mu_i - \mu_j)^2 + (\theta_i - \theta_j)^2} \right]^{1/2}, \quad i \neq j \quad (5.31)$$

such that the spectral norm of $[(-A_i^T) \oplus A_i]^{-1}$ decreases as

the distance between the eigenvalues of A_1 and A_2 increases. Now, notice that (5.22b) implies

$$\|P_y(k-1-1)\|_S \leq \|(-A_1^T)^{k-1-1} \oplus A_1^{k-1-1}\|_S \|(-A_1^T) \oplus A_1\|_S = (k-1) \quad (5.32)$$

It then follows from (5.31), (5.32), Lemma 5 and Theorem 2, that the maximum γ in (5.28) (i.e. the maximum uncertain coupling) for which stability can be assured is proportional to the distance between the eigenvalues of A_1 and A_2 . This phenomena, though intuitive, cannot be shown by the results of Lemma 4 and Theorem 1.

Remark 5. If the initial condition $x(0)$ is actually known (e.g. $x(0) = 0$), then the nominal state $x_0(k)$ is also known. In this case it is possible to substitute $x_0(k)$ for $\hat{x}_0(k)$ in Lemmas 4 and 5 to obtain better (i.e. smaller) perturbation bounds $\hat{e}^{(1)}(k)$ and $\hat{e}^{(2)}(k)$.

6. An example

Consider the uncertain system

$$z(k+1) = (A^0 + E)z(k) + w(k), \quad z(0) \quad (6.1)$$

$$v(k) = (c^0)^T z(k) \quad (6.2)$$

where $z \in \mathbb{R}^4$,

$$A^0 = \text{block-diag} \begin{bmatrix} -a_1 & -a_2 \\ 1 & 1 \end{bmatrix}, \quad a_i \in [1] \quad (6.3)$$

$$(c^0)^T = [1, -1, 0, 0] \quad (6.4)$$

and

$$w(k)^T = [1, -1, 0, 0] \quad (6.5)$$

The uncertainty matrix

$$E \in \mathbb{E} = \{E \in \mathbb{R}^{4 \times 4} : \|E\|_M = \gamma\} \quad (6.6)$$

where

$$\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix}$$

Let

$$T = \frac{1}{\sqrt{2}} \text{block-diag} \left\{ \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right\} \quad (6.8)$$

and make the substitution $z(k) = Tx(k)$. Then, the system described by (6.1)–(6.7) transforms to the scalar subsystem form

$$x(k+1) = (A + G)x(k) + v(k), \quad x(0) = 0 \quad (6.9)$$

where

$$A = T^{-1}A^0T = \text{diag} \{v_1 + j\omega_1, v_1 - j\omega_1, v_2 + j\omega_2, v_2 - j\omega_2\} \quad (6.10)$$

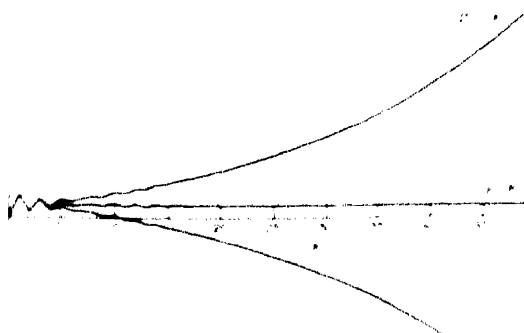
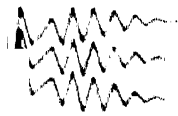


FIG. 1 Performance bounds from Theorem 1



Performance bounds from Theorem 2

$$c^T = (c^0)^T T = \frac{1}{\sqrt{2}} [1, -1, 1, 0, 0] \quad (6.11)$$

$$v(k)^T = [c^T w(k)]^T = \frac{1}{\sqrt{2}} [1, -1, 1, 0, 0], \quad k \geq 0 \quad (6.12)$$

and

$$G \in \mathbb{G} = \{G \in \mathbb{R}^{4 \times 4} : \|G\|_M = \gamma, G^T = G\} \quad (6.13)$$

where

$$G = (T^{-1}E^T + ET)_{\mathbb{M}} = \gamma \text{block-diag} \left\{ \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right\} \quad (6.14)$$

Now, let

$$v_1 = 0.8, \quad \omega_1 = 0.2, \quad v_2 = 0.5, \quad \omega_2 = 0.2, \quad \gamma = 0.05 \quad (6.15)$$

(Note that for these choices of v_i and ω_i the system is nominally stable.) Figures 1 and 2 then show the possible performance degradation in the nominal output as predicted respectively by Theorems 1 and 2 (with $G_1 = G$).

The M matrix condition of Theorem 1 predicts the instability seen in Fig. 1. Also notice that as guaranteed by the theory the performance bounds of Fig. 2 are less conservative than those of Fig. 1.

7. Conclusion

This paper has used majorant bounding techniques to develop time-dependent upper and lower performance bounds for a class of uncertain linear discrete time systems. The main theorems have also presented sufficient conditions for robust stability. The robustness results of Theorem 2 capture the intuitive effect that increasing the distance between the spectra of two subsystems increases the magnitude of nondestabilizing uncertain coupling allowed between the subsystems. A comparison of Theorems 2 and 3 also demonstrates that alternative representations of a given uncertain system can lead to significantly different robustness bounds.

Acknowledgements.—We thank Jill Strachla and Linda Ford for typing the original manuscript and Scott Greeley for performing the numerical calculations.

This work was supported in part by the Air Force Office of Scientific Research under contract F49620-86-C-0038.

References

- Berman, A. and R. J. Plemmons (1979) *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York.
- Brewer, J. W. (1978) Kronecker products and matrix calculus in systems theory. *IEEE Trans. Circ. Syst.*, **CAS-25**, 772–781. (Also, Correction to "Kronecker products and matrix calculus in systems theory," **CAS-26** (1979), 860).

Collins, E. G. Jr and D. C. Hyland (1989). Improved robust performance bounds in covariance majorant analysis. *Int. J. Control*, **50**, 1259-1266.

Dahlquist, G. (1983). On matrix majorants and minorants with applications to differential equations. *Lin. Alg. Appl.*, **52/53**, 199-216.

Fiedler, M. and V. Ptak (1962). On matrices with non-positive off-diagonal elements and positive principal minors. *Czech. Math. J.*, **12**, 382-400.

Hyland, D. C. and D. S. Bernstein (1987). The majorant-Lyapunov equation: A nonnegative matrix equation for robust stability and performance of large-scale systems. *IEEE Trans. Aut. Control*, **AC-32**, 1005-1013.

Hyland, D. C. and E. G. Collins, Jr (1989a). An M -matrix and majorant approach to robust stability and performance analysis for systems with structured uncertainty. *IEEE Trans. Aut. Control*, **34**, 691-710.

Hyland, D. C. and E. G. Collins, Jr (1989b). Block Kronecker products and block norm matrices in large-scale systems analysis. *SIAM J. Matrix Anal. Appl.*, **10**, 18-29.

Lasley, F. L. and A. N. Michel (1976). Input-output stability of interconnected systems. *IEEE Trans. Aut. Control*, **AC-21**, 84-89.

Ostrowski, A. M. (1961). On some metrical properties of operator matrices and matrices partitioned into blocks. *J. Math. Anal. Appl.*, **2**, 161-209.

Porter, D. W. and A. N. Michel (1974). Input-output stability of time varying nonlinear multiloop feedback systems. *IEEE Trans. Aut. Control*, **AC-19**, 422-427.

Seneta, E. (1973). *Non-Negative Matrices*, Wiley, New York.

Sivan, G. P. H. (1973). Hadamard products and multivariate statistical analysis. *Lin. Alg. Appl.*, **6**, 217-240.

Appendix A: Proof of Lemma 6

Recall (Brewer, 1978) that

(A ⊗ B)(C ⊗ D) = (AC) ⊗ BD (A.1)

It follows from (5.15) and (A.1)

℘_{ij}(k - l - 1) = (I_{n_i} ⊗ A_i^{k-l-2}) ∑_{s=0}^{k-l-2} (A_j^s ⊗ A_i^{s-1}) (A.2)

which for i = j gives (5.23a)

Now assume that i ≠ j. Then, since A_i and A_j do not have common eigenvalues, it follows (Brewer, 1978) that the n_in_j × n_in_j matrix (A_j^s ⊗ A_i^{s-1}) does not have any eigenvalue equal to one. Thus, (I_{n_{n_{ij}}} - A_j^s ⊗ A_i^{s-1}) is nonsingular. It then follows from (2.16) and (A.2) that

℘_{ij}(k - l - 1) = (I_{n_i} ⊗ A_i^{k-l-2}) [I_{n_{n_{ij}}} - (A_j^s ⊗ A_i^{s-1})^{k-l-1}] × [I_{n_{n_{ij}}} - (A_j^s ⊗ A_i^{s-1})]⁻¹, i ≠ j (A.3)

Using (A.1) and the fact that I_{n_{n_{ij}}} = I_{n_i} ⊗ I_{n_j} it follows from (A.3) that

℘_{ij}(k - l - 1) = [I_{n_i} ⊗ A_i^{k-l-2} - (A_j^s)^{k-l-1} ⊗ A_i^{s-1}] × [I_{n_{n_{ij}}} - (A_j^s ⊗ A_i^{s-1})]⁻¹, i ≠ j (A.4)

Now since A_j is nonsingular, it follows (Brewer, 1978) that (I_{n_j} ⊗ A_j) is also nonsingular. Thus

℘_{ij}(k - l - 1) = [I_{n_i} ⊗ A_i^{k-l-2} - (A_j^s)^{k-l-1} ⊗ A_i^{s-1}] × (I_{n_i} ⊗ A_j)(I_{n_i} ⊗ A_j)⁻¹ [I_{n_{n_{ij}}} - (A_j^s ⊗ A_i^{s-1})]⁻¹, i ≠ j (A.5)

which using (A.1) becomes

℘_{ij}(k - l - 1) = [I_{n_i} ⊗ A_i^{k-l-1} - (A_j^s)^{k-l-1} ⊗ I_{n_i}] × [I_{n_i} ⊗ A_j - A_j^s ⊗ I_{n_i}]⁻¹, i ≠ j (A.6)

The proof is completed by simply noting that (A.6) and (5.22b) are equivalent. □

Appendix B: Proof of Theorem 3

First note that (5.17) is equivalent to

ê⁽¹⁾(k) = ∑_{l=0}^{k-1} A_s^{k-l-1} Ĝ_r [ê⁽¹⁾(l) + x⁽⁰⁾(l)], ê⁽¹⁾(0) = 0 (B.1)

Substituting (B.1) into its own right hand side, utilizing (4.16) and exchanging l for m and m for l gives

ê⁽¹⁾(k) = ∑_{l=0}^{k-2} (℘(k - l - 1) * Ĝ_r) Ĝ_r [ê⁽¹⁾(l) + x⁽⁰⁾(l)] + ∑_{l=0}^{k-1} (A_s)^{k-l-1} Ĝ_r x⁽⁰⁾(l), ê⁽¹⁾(0) = 0 (B.2)

where ℘(k - l - 1) ∈ ℝ^{r × r} is given by

℘_{ij}(k - l - 1) = ∑_{s=0}^{k-l-2} ||A_j||_s^{k-l-1-2} ||A_j||_s², i, j = 1, ..., r (B.3)

It follows from (B.3), (5.15) and (2.15) that

||℘_{ij}(k - l - 1)||_s ≤ ∑_{s=0}^{k-l-2} ||(A_j^s) ⊗ A_i^{s-1-2}||_s = ℘_{ij}(k - l - 1), i, j = 1, ..., r (B.4)

or equivalently

℘_{ij}(k - l - 1) ≤ ℘(k - l - 1). (B.5)

Then from (B.5) and (5.17) it follows that

ê⁽²⁾(k) ≤ ∑_{l=0}^{k-2} [℘(k - l - 1) * Ĝ_r] Ĝ_r [ê⁽²⁾(l) + x⁽⁰⁾(l)] + ∑_{l=0}^{k-1} (A_s)^{k-l-1} Ĝ_r x⁽⁰⁾(l), ê⁽²⁾(0) = 0. (B.6)

The inequality (5.25) follows from (B.6) and Lemma 3.

The inequalities (5.26) and (5.27) are an immediate result of (5.26) and the definitions of α⁽¹⁾(k), β⁽¹⁾(k), α⁽²⁾(k) and β⁽²⁾(k) given respectively by (5.11), (5.12), (5.23) and (5.24).

Recall that the robust stability condition (5.13) holds if and only if for any x⁽⁰⁾(·) which is a nominal initial condition response for a strictly nonzero initial condition lim_{k → ∞} ê⁽¹⁾(k) = 0. That the robust stability condition (5.13) is a sufficient condition for the robust stability condition of Theorem 3 is thus a direct result of (5.25). □

Brief Paper

Rapid Tracking of Complex Trajectories in Short-duration Processes*

C. C. H. MA†

Key Words—Tracking systems, discrete-time systems, feedforward control, saturation control, repetitive control

Abstract—A feedforward methodology for rapid tracking of complex reference trajectories in short-duration linear industrial processes is proposed. The methodology allows easy bounding of the actuation signal magnitude. It is basically iterative in nature, but relatively fast convergence can be achieved in general with a compact formulation of the iteration algorithm, and particularly fast when the system to which the algorithm is applied has fast dynamics. Simulation results indicate that the methodology has properties desirable in a practical implementation.

1. Introduction

MAKING THE output of a machine track a given reference trajectory is a common industrial problem. For example, to pick and place an object, mill a precise pattern out of a rough workpiece, assemble and disassemble equipment etc. all require the machines involved to track a reference trajectory. For some of these problems, such as pick and place, only the final position and/or orientation of the machine may be of importance and the in-motion tracking performance is often of little value. For these problems, time can usually be traded off for better tracking accuracy, or the trajectories may be compromisingly preplanned to lower their orders to accommodate the plant with a low-order controller (Kim and Shin, 1985; Liu and Chang, 1983). However, if a task requires a machine to track the reference trajectory accurately during most part of the (usually short duration) task, such as in assembly-disassembly, or for increased productivity, the timing of the motion of the machine is important. The ability to rapidly track the short reference trajectory by the machine then becomes highly desirable.

Given a linear time-invariant plant [e.g. a globally feedback-linearized plant (Bertoz, 1974; Freund, 1982) or a Cartesian robot] connected to a linear controller in the standard feedback control configuration and given an arbitrary reference trajectory with transfer function R , it is known (Francis and Vidyasagar, 1983) that unless the loop transfer function is divisible by the largest invariant factor of the “denominator” of R , the autonomous feedback control system cannot track the reference trajectory, even asymptotically. If the reference trajectory is complex so that R is of high order, the controller would have to be of high order as well in order to achieve just asymptotic tracking. Although delayed asymptotic tracking (i.e. tracking with some constant delay) may be accomplished with a deadbeat controller and may be satisfactory for many applications,

such a controller can be designed only for stably invertible plants. In any case, it is only *asymptotic* tracking of the reference that an autonomous linear time-invariant feedback controlled system can achieve, at best. For the second class of tasks mentioned previously, where short yet possibly highly complex reference trajectories need to be precisely tracked, preferably instantly (if practically possible) but definitely comparatively long before the end of the task, a different control methodology must be applied.

In this paper, we propose a control methodology which may enable a linear discrete-time plant to track any short reference trajectory precisely, quickly and possibly instantly, and to take into account possible constraints on the control signal. The basic methodology is iterative in nature, but it has been formulated such that iterations may be carried out exponentially fast with respect to the computational time requirement.

2. Rapid tracking system structure

Figure 1 shows a basic control system structure for rapid tracking by the plant P , where r denotes the reference trajectory to be tracked and y denotes information other than r which is essential to the generation of the control signal. Unless y is a function of x in real-time operation, it is well known that such open-loop system is highly sensitive to external noise. To minimize this sensitivity, as well as to stabilize the plant in case it is open-loop unstable, a stabilizing feedback controller C can be added (Francis and James, 1984; Chang and Pearson, 1984; Vidyasagar, 1985), resulting in the proposed tracking system structure shown in Fig. 2. The signal q shall be referred to as the tracking control.

It is easy to verify that since C stabilizes the feedback loop, when precision tracking is taking place, the magnitude of x will be close to zero and u will be close to q . Therefore, by bounding q , the control signal u can be kept small to facilitate implementation of the tracking system.

3. Instant tracking control

Referring to Fig. 2, let e^k and $\{e_i^k\}_{i=1}^N$ denote the tracking error and its sequence of N values corresponding to time $k = 1, \dots, N$ respectively when the tracking control is $q^k = \{q_i^k\}_{i=1}^N$. Let $h = \{h_i\}_{i=1}^N$ denote the strictly causal impulse response sequence at output y due to a unit impulse applied at $k = 0$ to input q , and let Δq be the modification to be made on q to reduce the current error energy $J^k = (e^k)^T e^k$, even further. Since, for a linear time-invariant system, the input signal causing an output equals the deconvolution of the output signal with the system impulse response, the optimal Δq required for nullifying J^k must satisfy

$$e^k = h \otimes \Delta q^{k+1} + H \Delta q^{k+1}, \quad \otimes = \text{convolution},$$

$$H = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix}$$

* Received March 24, 1988; revised 19 May 1989; revised 20 March 1990; received in final form 25 April 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernack.

† Department of Electrical Engineering, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5.

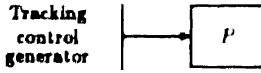


FIG. 1 Basic tracking system structure.

Hence

$$\Delta q^{(n)} \quad (1)$$

would uniquely reduce $J^{(n)}$ to zero provided that $h_1 \neq 0$. This can also be obtained by minimizing the tracking error energy $J^{(n)} \Delta q$ with respect to Δq . We shall assume $h_1 \neq 0$ in this paper unless otherwise stated. This can be done without loss of generality (via re-indexing the impulse responses if necessary).

Although the optimal tracking control modification in (1) can theoretically reduce $J^{(n)}$ to zero, it is easy to see that if $|h_1|$ approaches zero, the magnitude of Δq , and hence that of the resulting control $q^{(n)} = q^{(n-1)} + \Delta q^{(n)}$, will be unbounded and implementation of $q^{(n)}$ will not be possible since there will always be constraints on the control signal magnitude in practice.

The instant exact tracking that $\Delta q^{(n)}$ can theoretically provide is perhaps more than needed for most practical purposes however. This is because most of the industrial tasks under consideration seem to be characterizable by three stages of operation: (1) approaching, (2) performing and (3) settling. In the approaching stage, the machine either approaches the work-piece or simply approaches the work-area and trajectory tracking need not be very precise at the beginning. In the performing stage, the work-piece is actually being worked on or maneuvered and high precision tracking is desired. Finally in the settling stage, the machine leaves the work-piece or -area to settle into a final still position (most likely also the initial position), ready for the next task. As for the approaching stage, more tracking error can be tolerated as the machine moves farther away from the work-piece or -area in the settling stage. However, one good feature for the control system is that the tracking control signal converges to small (if not zero) value before the end of the tracking task. This is desirable so that the plant need not be driven to the final still position *dynamically*. (If the plant needs to be driven dynamically to the final still position, it will need limit switch(es) to keep it from wandering when the control signal is removed.) The following methodology has just such properties.

4. Proposed tracking control

Suppose instead of modifying the entire tracking control vector $q^{(n)}$ at once to nullify the resulting tracking error, we modify the individual elements in $q^{(n)}$ one at a time, minimizing the resulting error energy each time. Suppose further we modify them in the sequence $q_0, q_1, q_2, \dots, q_{N-1}$, then cycle until tolerable error energy is obtained. This results in a tracking control generation (or *q-learning*) methodology which would normally allow a particular variation to take place on an element of q only if such variation will lead to the maximum possible decrease in the resulting tracking error energy. Should such variation result in the particular control value exceeding the bound set on q , the variation can be simply limited to meet the bound. Then the learning process can be allowed to continue, if more reduction on the error energy is desired.

To formulate the optimal modification to make on the k th control element in order to minimize the resulting error energy, suppose l cycles of modifications had been made. Let $q^{l,k}$ denote the current tracking control (i.e. the control obtained after the $k-1$ th element of q has been modified),

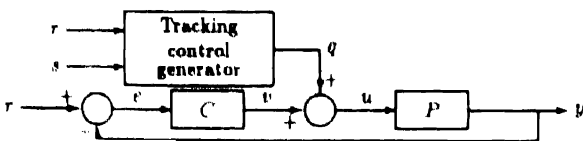


FIG. 2 Tracking system structure.

$e^{l,k-1}$ denote the corresponding tracking error and $\Delta q_k^{l,k-1}$ denote the modification to be made on the k th element of q . (With notations defined as such, it is clear that $q_k^{l-1} = q_k^{l-1,N-1}$, $e^{l-1} = e^{l-1,N-1}$ and $\Delta q_0^{l-1} = \Delta q_0^{l-1,N-1}$, q^{l-1} and e^{l-1} may also be simply denoted by q^l and e^l respectively.) Then the control and error energy after $\Delta q_k^{l,k-1}$ is added to $q^{l,k-1}$ will be

$$q_k^{l,k} = q_k^{l,k-1} + \Delta q_k^{l,k-1}, \quad e^{l,k} = e^{l,k-1} - \Delta q_k^{l,k-1} h^k, \quad (2)$$

where $h^k = [0 \dots 0 h_1 h_2 \dots h_{N-k}]^T$ with k leading zeros. If $\Delta q_k^{l,k-1}$ were to minimize $J^{l,k} = (e^{l,k}, e^{l,k})$, it must satisfy

$$\frac{dJ^{l,k}}{d\Delta q_k^{l,k-1}} = 0 = -2h^k (e^l - \Delta q_k^{l,k-1} h^k). \quad (3)$$

This yields the optimal k th modification

$$\Delta q_k^l = \frac{(h^k, e^{l,k-1})}{(h^k, h^k)}, \quad k = 0, \dots, N-1, \quad (4)$$

and substitution into equations (2) results in

$$e^{l,k} = e^{l,k-1} - \Delta q_k^{l,k-1} h^k = \frac{h^k (h^k, e^{l,k-1})}{(h^k, h^k)} \quad (5)$$

Equations (4)–(5) shall be referred to as the *q-learning Algorithm 1* from now on. With $q^{l,k}$ and $e^{l,k}$ now available, the process can go on to calculate the next optimal modification $\Delta q_{k+1}^{l,k}$ to yield $q^{l,k+1}$ and $e^{l,k+1}$.

Lemma 1. The tracking control *q-learning Algorithm 1* made up by equations (4)–(5) converges with $e^{l,k} \rightarrow 0$ and $q^{l,k} \rightarrow q^{(n)}$ as the iteration continues.

Proof. Since the solution in equation (4) is a minimum point, $J^{l,k}$ is non-increasing. Since $J^{l,k}$ is also lower bounded, it must converge. Equation (4) is the unique minimum solution, therefore $J^{l,k}$ cannot stay unchanged unless $\Delta q_k^{l,k-1} = 0 \forall k$. This implies that $e^{l,k}$ must converge as $J^{l,k}$ does. Now suppose $e^{l,k}$ converges to e^* (which need not equal zero at this point of the proof). Then $\Delta q_k^{l,k-1} = 0 \forall k$ together with equation (4) imply that $(h^k, e^*) = 0 \forall k$. Since $(h^k, e^*) = H^l e^*$ and H is nonsingular, they further imply that $e^* = 0$. It was shown in Section 3 that zero tracking error can be achieved uniquely by $q^{(n)}$ only. Therefore, as $e^{l,k}$ converges to zero, $q^{l,k}$ must converge to $q^{(n)}$. \square

Algorithm 1 can be very computation intensive when N and/or l is large. Fortunately, it can be shown that the formulation for q^{l+1} and e^{l+1} can be compacted to take the form

$$e^l = (I - H \Sigma^{-1} H^l) e^{l-1} = (I - H \Sigma^{-1} H^l) e^0, \quad (6)$$

$$q^l = H^{-1} (e^0 - e^l) + q^0 = H^{-1} (I - (I - H \Sigma^{-1} H^l)^l) e^0 + q^0, \quad (7)$$

$$\Sigma = \sum_{k=0}^{N-1} h_k h_k^T$$

$$\Sigma = \sum_{k=0}^{N-1} h_k h_k^T = \sum_{k=0}^{N-1} h_k h_k^T$$

$$h_k h_k^T = \sum_{k=0}^{N-1} h_k h_k^T, \dots, \sum_{k=0}^{N-1} h_k h_k^T$$

Equations (7) and (6) together shall be referred to as the *q-learning Algorithm 2*.

Algorithm 2 is much more computationally efficient for large l (more accurately, for approximately $l > (N+2) \log_2 l$) than Algorithm 1. This is clear since for any square matrix T and for l equal to some power of 2, T^l can be computed as $(\dots ((T^2)^2) \dots)^2$ in $\log_2 l$ number of matrix multiplications. This efficiency will become even more pronounced with the advent of the fast array processors.

Algorithm 2 can be used for as long as the resulting

q^i does not exceed the possible magnitude bound set on its values. Suppose q^i does not exceed the bound but q^{i+1} does, and more reduction on the tracking error energy $J^i = (e^i, e^i)$ is desired. If l is still small, the best approach is to switch from using Algorithm 2 to Algorithm 1 after q^i and e^i are obtained and continue with the q -learning process. If l is large, then since Algorithm 2 is more computationally efficient, it may be better to keep using Algorithm 2 for as long as possible before switching to use Algorithm 1. In that case, the well-known binary-search technique may be followed to find the best q which does not exceed the magnitude bound.

The result of these q -learning procedures will be bounded tracking control q that, when applied to the system, will yield a tracking error e with small energy. Due to the fact that e_k in $e = (e_k)_1^N$ gets taken into account k number of times each modification cycle in the process, the resulting value of e_k tends to be larger (in magnitude) when k is small than when k is large. Hence the tracking performance will be comparatively the worst at the beginning, then becomes better as time elapses.

5. Convergence rate of Algorithm 1

Equation (6) implies that when Algorithm 1 is applied, the tracking error energy will decrease over the $l + 1$ th learning-cycle by the amount

$$\begin{aligned} \|e^l\|^2 - \|e^{l+1}\|^2 &= (e^l)'[I - (I - H\Sigma^{-1}H')^l(I - H\Sigma^{-1}H')]e^l \\ &= (e^l)'[H(\Sigma^{-1})'[\Sigma' + \Sigma - H'H]\Sigma^{-1}H']e^l = (e^l)'Te^l \\ &\geq \lambda_{\min}(T)\|e^l\|^2 \end{aligned}$$

After some manipulations it can be shown that

$$T = \Gamma - H(\Sigma^{-1})' \begin{bmatrix} \sum_{i=1}^N h_i h_i' & & \\ & \sum_{i=1}^{N-1} h_i h_i' & 0 \\ 0 & & \sum_{i=1}^1 h_i h_i' \end{bmatrix} - \Sigma^{-1}H' > 0$$

Therefore, the closer λ_{\min} is to 1 the faster the rate of convergence. From the forms of H and Σ it can be seen that if $h_1 \neq 0$ while $h_k = 0 \forall k \geq 2$, then $T = I$, $\lambda_{\min} = 1$ and convergence will take place in must *one* single learning cycle. For any other impulse response sequence the convergence rate will be slower.

The value λ_{\min} represents the worst case rate of convergence of Algorithm 1 when applied to a given system. It is relatively difficult to compute in general, especially when N is large. Also in practice the actual convergence rate may be little or much faster than λ_{\min} depending on how close to 1 the remaining eigenvalues of T are, as well as on the tracking error sequence to be eliminated. For these reasons, we suggest using the product of the eigenvalues of T for comparisons of convergence rates when applying Algorithm 1 to different systems with roughly the same λ_{\min} s. Since the product of eigenvalues of a constant matrix equals the determinant of the matrix, the product Λ of eigenvalues of T is simply

$$\Lambda = h_1^{2N} \prod_{i=1}^N \left(\frac{1}{\prod_{j=1}^i h_j^2} \right) \prod_{j=1}^N \left(\frac{1}{\prod_{i=1}^j (h_i^2)} \right) \quad (8)$$

The above equation indicates that the larger the value of h_1 relative to the rest of the impulse response sequence the closer to 1 Λ will be. This implies that Algorithm 1 will converge fast for a system with fast dynamics between input q and output y ; in other words, a system whose impulse response both rises to its peak value then decreases to zero very quickly. This is consistent with the observation that convergence will take place in just one learning cycle if $h_k = 0$ for all k except 1. Therefore, in order to promote fast

convergence of Algorithm 1, it is desirable to design C so that the stable feedback system poles are close to the origin of the z -plane.

6. Some application-related issues on the methodology

The q -learning process proposed assumes the availability of q^0 and e^0 . If $q^0 = 0$, then e^0 is the nominal tracking error before the application of any tracking control. Such e^0 would normally be available in practice only if the tracking system is used for repetitive tracking of the same reference trajectory, such as may be found in manufacturing. In this case, the proposed tracking control methodology can be used for on-line improvement of the tracking performance (or self-learning).

Suppose the nominal tracking error of the system is not available. Then note that the system in Fig. 2 is equivalent to the system in Fig. 3. If this equivalent system is initially relaxed, then w (instead of q) can be generated using the q -learning process assuming $e^0 = 0$ and w^0 (instead of q^0) $= 0$. Then q can be obtained by subtracting z from w . Note that the tracking control q was designed to be injected into the system after the controller C (see Fig. 2) mainly to facilitate the bounding of u (by bounding q). If the controller is a simple gain, say K , then bounding u becomes an easy job whether q is injected after or before the controller.

The q -learning methodology proposed also assumes the availability of $\{h_k\}_1^N$, which is the impulse response at y due to an impulse applied at input q . Such response can actually be identified on-line for the system in Fig. 2 by using data sequences $\{w_k\}_0^{N-1}$ and $\{y_k\}_1^N$. This is clear since

$$y = w \otimes h, \quad \begin{bmatrix} w_0 & & & 0 \\ w_1 & w_0 & & \\ & \ddots & \ddots & \\ w_{N-1} & & w_1 & w_0 \end{bmatrix} h,$$

if the system is initially relaxed. Any popular one-shot or recursive identification algorithm can be used. In case the system were not initially relaxed when $\{w_k\}_0^{N-1}$ and $\{y_k\}_1^N$ were collected, it may be better to identify the (possibly) fewer parameters of the transfer function from w to y first, then derive the impulse response $\{h_k\}_1^N$ from the resulting transfer function. This is because identifying the parameters of the transfer function requires much fewer accurate sets of data, and the last few data in $\{y_k\}_1^N$ reflect the impulse response of the system more accurately since they are not affected as much by the initial conditions of the system as the first few data are.

Although convergence of the q -learning process will be theoretically slow when h_1 is small relative to the rest of the impulse response sequence, it may still be speeded up in practice. This may be done, for example, by assuming that the first few small values of the impulse response to be zero (i.e. ignoring them) and reindexing the impulse response so that h_1 has a relatively large magnitude. Next one should iterate to obtain q until the tracking error energy is relatively small (this is a rather crude idea), then switch back to the original impulse response sequence and continue iterating until satisfactory error energy is obtained.

7. Simulation results

Many single-input single-output second order systems were simulated in the configuration of Fig. 2. One of them has the

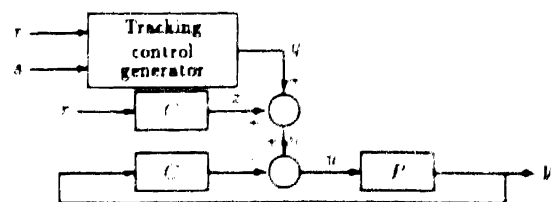


FIG. 3. Equivalent tracking system

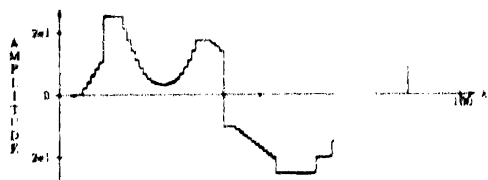


FIG. 4. Plot of r and y

dynamics

$$\begin{aligned} &1.155y(k) - 1.0y(k - 1) - 1.5y(k - 2) \\ &\quad = -1.15u(k - 1) + 2.5u(k - 2), \\ &v(k) + 4.18v(k - 1) = 3.503e(k) + 2.849e(k - 1), \\ &y(k) = 0.0 \quad \text{for } k \leq 0. \end{aligned}$$

The proposed q -learning algorithm was applied to compute the tracking control assuming e^0 is available (not necessary). Figures 4 and 5 show the results of the system after applying Algorithm 1 to learn q for 25 learning cycles.

Note that the plant in this system is neither stable nor minimum-phase. If the q -learning process were allowed to continue for another five cycles, the tracking error would become practically undetectable. By setting the magnitude bound on q to be 22.0, fairly precise tracking is still obtained after only 25 q -learning cycles. The results are shown in Fig. 6. Of course, since q has been bounded below that which is required for perfect tracking, perfect tracking cannot be achieved even if the learning process were allowed to go on forever.

In all of the simulation results, the tracking control signal always seems to be trailed by small (if not zero) values. This is characteristic of the control methodology proposed. It is due to modifying the tracking control values in the sequence $q(0), q(1), \dots, q(N)$. By the time the last $q(k)$ s are modified, the tracking error energy is already reduced significantly. What remains for the last $q(k)$ s to reduce is small, hence resulting in small trailing tracking control signal values. Such a property is very desirable in practice.

8. Conclusions

A feedback control methodology has been proposed for the practical precise tracking of short reference trajectories of orders significantly higher than that of the plant. The controlled system consists of a standard feedback control loop with an additional input from which an auxiliary "tracking control" signal is injected. The feedback controller stabilizes the plant in case it is open-loop unstable as



FIG. 5. Plot of q and v

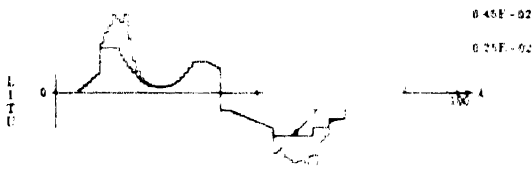


FIG. 6. Plot of r and y with q bounded.

well as minimizing the noise sensitivity of the system, while the tracking control minimizes the tracking error in the least squares sense.

The generation of the tracking control is iterative in nature, but a compact formulation has been developed such that the iterations can be carried out exponentially fast. Besides being able to make the plant track the reference trajectory more accurately, it turns out that such methodology will generally yield a tracking control signal which converges to a small value before the end of the short tracking task. This will make it easier for the plant to settle into its final still position at the end of the task.

The iteratively generated tracking control will converge asymptotically to the "instant tracking control" for the system. However, the rate of convergence depends on the magnitude of the first impulse response sample relative to those of the rest. Hence one can expect in general that the faster the dynamics of the feedback loop the faster the convergence will take place.

Acknowledgements—This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant OGP-0036382 and by the University of British Columbia under the equipment grant 5-80289.

References

Bejczy, A. C. (1974) Robot arm dynamics and control. *JPL Technical Memo*, pp. 33-369.

Chang, B. C. and J. B. Pearson (1984) Optimal disturbance reduction in linear multi-variable systems. *IEEE Trans. Aut. Control*, **AC-29**, 880-887.

Francis, B. A. and M. Vidyasagar (1983) Algebraic and topological aspects of the regulator problem for lumped linear systems. *Automatica*, **19**, 87-90.

Francis, B. A. and G. Zames (1984) On H^∞ -optimal sensitivity theory for SISO feedback systems. *IEEE Trans. Aut. Control*, **AC-29**, 9-16.

Freund, E. (1982) Fast nonlinear control with arbitrary pole-placement for industrial robots and manipulators. *Int. J. Robotics Res.*, **1**, 65-78.

Kim, B. K. and K. G. Shin (1985) Minimum-time path planning for robot arms and their dynamics. *IEEE Trans. Syst. Man Cybern.*, **SMC-15**, 213-223.

Lin, C.-S. and P.-R. Chang (1983) Joint trajectories of mechanical manipulators for cartesian path approximation. *IEEE Trans. Syst. Man Cybern.*, **SMC-13**.

Vidyasagar, M. (1985). *Control System Synthesis: A Factorization Approach*. MIT Press, Cambridge, MA, 1985.

Brief Paper

Optimal Hold Functions for Sampled Data Regulation*

Y.-C. JUAN† and P. T. KABAMBA‡‡

Key Words—Digital control, feedback control, optimal regulators, sampled data systems

Abstract—This paper investigates the use of Generalized Sampled Data Hold Function Control (GSHF) to optimize quadratic measures of performance in sampled data control loops. The idea of GSHF control is to use sampled data feedback, but consider the hold function as a design parameter. Explicit solutions are given for the GSHF versions of the optimal LQ and LQG regulators. The questions of existence, uniqueness and computation of optimal hold functions are treated. An example is presented.

1. Introduction

THE TRADITIONAL approach to sampled data control assumes that the discrete time plant model is obtained from a continuous time system by using prespecified hold devices—typically, zero-order or first-order hold (Franklin and Powell, 1980). A digital computer then implements a discrete time control law. In such a scheme, the subsequent intersampling behavior of the controlled continuous time system is fairly well understood (De Souza and Goodwin, 1984; Berger, 1985; Urkura and Nagata, 1987; Sirisena, 1988; Franklin and Emami-Naeini, 1986). Recently, however, a new approach to sampled data control has been introduced (Chammas and Leondes, 1978a, b, 1979). This method is called "Generalized Sampled Data Hold Function Control" (GSHF), and its original feature is to consider the hold function as a design parameter. Until now, studies on GSHF control have been primarily focused on the resulting discrete time system, and little has been said about the intersampling behavior of the controlled continuous time system (Juan and Kabamba, 1987; Kabamba, 1986, 1987; Kaczorek, 1985, 1986; Zaygren, 1983; Zeng, 1985; Zaygren and Tarn, 1984; Tarn *et al.*, 1988; Greshak and Verghese, 1982).

In this brief paper, we present solutions of optimal design problems for GSHF control where the performance index penalizes the intersampling behavior of the closed loop system. This is accomplished by using penalty indices which are time integrals of a function of the state and control of the regulated continuous time system. The optimization problem then becomes a standard optimal control problem, but where the "control variable" is the hold function itself. Existence and uniqueness of an optimal hold function is proved for a wide class of optimal GSHF regulation problems. The specific problems solved in this paper are GSHF-control versions of the linear-quadratic and linear-quadratic Gaussian regulators. The solution of these problems, together with our results on existence and uniqueness of optimal hold functions constitute the original contribution of this study.

* Received 31 May 1988; revised 7 March 1990; received in final form 23 July 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak. Partially supported by NSF under Grant ECS-8722244.

† Aerospace Engineering Department, The University of Michigan, Ann Arbor, MI 48109-2140, U.S.A.

‡ Author to whom all correspondence should be addressed.

Several authors have investigated the intersampling behavior of sampled data control systems. Dorato and Lewis (1971) showed how to compute the penalty matrices of a discretized LQG problem based on those of a continuous time LQG problem, assuming zero order hold. De Souza and Goodwin (1984) presented a method for computing the time varying output covariance between sampling instants. Berger (1985) suggested that the output between samples can be predicted by a linear function and used this in adaptive control. Ripple free deadbeat control for sampled data systems was investigated in Urkura and Nagata (1987) and Sirisena (1988), and a necessary and sufficient condition was given. Franklin and Emami-Naeini (1986) showed that a continuous time internal model is necessary and sufficient to provide ripple free response for the robust servomechanism. However, all these studies assume that the hold functions are not part of the design problem.

Generalized sampled data hold function control has been investigated for finite dimensional, continuous time systems (Kabamba, 1987; Juan and Kabamba, 1987; Chammas and Leondes, 1978a, b; Kabamba, 1986), finite dimensional discrete time systems (Kaczorek, 1985, 1986), and infinite dimensional, continuous time systems (Zaygren, 1983; Zeng, 1985; Zaygren and Tarn, 1984; Tarn *et al.*, 1988). The kindred idea of periodic compensation of time invariant systems has also received attention (Greshak and Verghese, 1982; Khargonekar *et al.*, 1985). The primary concern of these studies has been the properties of the discrete time closed loop system, because they determine important characteristics of the continuous time system such as stability and deadbeat response. As a consequence, the intersampling behavior of continuous time systems under GSHF regulation has received little attention. However, in Kabamba (1987) it was shown that in GSHF control, this intersampling behavior may be unsatisfactory. This phenomenon motivates the present study.

2. Problem formulation, notations and definitions

We use the following standard notations: superscript T denotes matrix transpose, $E(\cdot)$ denotes expected value, $\text{tr}(\cdot)$ denotes the trace of a matrix, $\delta(\cdot)$ denotes the Kronecker symbol in both the discrete time and continuous time case, I_p denotes the identity matrix of order p . Let $X \in \mathbb{R}^{m \times n}$, and denote the columns of X as $x_i, i = 1, \dots, n$, i.e. $X = [x_1 \dots x_n]$. The vec operator on X is defined as $\text{vec}(X) = \text{col}[x_1, x_2, \dots, x_n]$, where $\text{vec}(X) \in \mathbb{R}^{mn \times 1}$.

We consider finite dimensional linear time invariant, continuous time systems under sampled-data regulation as follows (see Fig. 1).

Plant and sampler

$$\dot{x}(t) = Ax(t) + Bu(t) + w(t) \quad (2.1)$$

$$z(k) = Cx(kT) + v(k) \quad (2.2)$$

Digital compensator

$$\xi(k+1) = P\xi(k) + U\zeta(k) \quad (2.3)$$

$$y(k) = S\xi(k) + V\zeta(k) \quad (2.4)$$

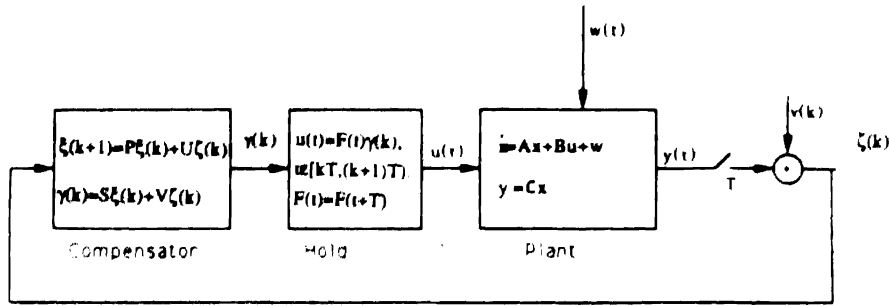


FIG. 1. Feedback configuration.

Digital-to-analog conversion (Hold device)

$$u(t) = F(t)\gamma(k), \quad t \in [kT, (k+1)T] \quad (2.5)$$

$$F(t) = F(t+T), \quad \forall t \quad (2.6)$$

where $x(t) \in \mathbb{R}^n$ is the plant state vector; $u(t) \in \mathbb{R}^m$ is the control input; $w(t) \in \mathbb{R}^p$ is a disturbance vector, $\xi(k) \in \mathbb{R}^p$ and $v(k) \in \mathbb{R}^p$ are the discrete measurement vector and discrete measurement noise vector, respectively; $\xi(k) \in \mathbb{R}^q$ is the compensator state vector; $\gamma(k) \in \mathbb{R}^r$ is the compensator output; $T > 0$ is the sampling period, $F(t) \in \mathbb{R}^{r \times q}$ is a T -periodic integrable and bounded matrix representing a hold function; and the real matrices A, B, C, P, U, S, V have appropriate dimensions. Without loss of generality, we assume that the matrices C and (S, V) have linearly independent rows.

The formalism of equations (2.1)–(2.6) is quite general. Zero-order hold (first-order hold, i th-order hold) control is obtained by letting the hold function $F(t)$ be a constant (first-degree polynomial, i th-degree polynomial, respectively). Also, by letting $q = 0$, $r = p$, and $V = I_p$, we obtain the formalism of GSHF control with direct output feedback used e.g. in Kabamba (1987).

For a given hold function $F(t)$, $t \in [0, T]$, the problem of designing the matrices P, U, S, V for performance of the corresponding discrete time system has been extensively treated in the literature [see e.g. Franklin and Powell (1980)]. Our objective in this paper is, for a given compensator (2.3)–(2.4), to determine time histories of the hold function $F(t)$ in (2.5), (2.6) that will optimize various performance criteria associated with the sampled data system (2.1)–(2.6).

Upon loop closure, the state and control between samples satisfy

$$\begin{aligned} x(kT+t) &= \Phi(t)x(kT) + D(t)S\xi(k) + D(t)Vv(k) \\ &\quad + \omega(kT+t), \quad t \in [0, T] \end{aligned} \quad (2.7)$$

$$\begin{aligned} u(kT+t) &= F(t)VC^{-1}(kT)x(kT) + F(t)S\xi(k) \\ &\quad + F(t)Vv(k), \quad t \in [0, T] \end{aligned} \quad (2.8)$$

where

$$\Phi(t) = \exp(At) + D(t)C, \quad t \in [0, T] \quad (2.9)$$

$$D(t) = \int_0^t \exp(A(t-\tau))BF(\tau) d\tau, \quad t \in [0, T] \quad (2.10)$$

$$\omega(kT+t) = \int_{kT}^{kT+t} \exp(A(kT+t-\tau))w(\tau) d\tau, \quad t \in [0, T] \quad (2.11)$$

Defining

$$x_a(k) = [x^T(kT), \xi^T(k)]^T \in \mathbb{R}^{n+q} \quad (2.12)$$

$$\Psi_a = \begin{bmatrix} \exp(AT) + D(T)VC^{-1} & D(T)S \\ UC^{-1} & P \end{bmatrix} \in \mathbb{R}^{(n+q) \times (n+q)} \quad (2.13)$$

$$D_a = \begin{bmatrix} D(T)V \\ U \end{bmatrix} \in \mathbb{R}^{(n+q) \times r} \quad (2.14)$$

$$\omega_a(k) = \begin{bmatrix} \omega(k) \\ 0 \end{bmatrix} \in \mathbb{R}^{n+p} \quad (2.15)$$

then the closed loop equations for the discrete time system are

$$x_a(k+1) = \Psi_a x_a(k) + D_a v(k) + \omega_a(k) \quad (2.16)$$

The closed loop monodromy matrix is defined as Ψ_a in (2.13), (2.16) and denotes the state transition matrix of the regulated discrete-time system over one period.

Definition. The design problem of finding an optimal hold function $F(\tau)$, $\tau \in [0, T]$ is called a *fixed monodromy (free monodromy)* problem if $D(T)$ in (2.10) is specified (not specified).

A fixed monodromy problem must therefore satisfy a design constraint of the form

$$D(T) = G, \quad (2.17)$$

where typically, the matrix G is chosen such that the closed loop monodromy matrix Ψ_a of (2.13) defines a stable discrete time system (2.16).

3 Linear-quadratic Gaussian regulation

Throughout this section we assume that $w(t)$ and $v(k)$ of (2.1)–(2.3) are stationary Gaussian processes satisfying

$$E[w(t)] = 0, \quad t \in \mathbb{R}, \quad (3.1)$$

$$E[v(k)] = 0, \quad k \in \mathbb{N}; \quad (3.2)$$

$$E[w(t)v^T(k)] = 0, \quad t \in \mathbb{R}, \quad k \in \mathbb{N}; \quad (3.3)$$

$$E[w(t)w^T(\tau)] = R_w \delta(t-\tau), \quad t, \tau \in \mathbb{R}, \quad R_w \geq 0; \quad (3.4)$$

$$E[v(k)v^T(i)] = R_v \delta(k-i), \quad k, i \in \mathbb{N}, \quad R_v \geq 0; \quad (3.5)$$

$$E[w(t)x^T(0)] = 0, \quad t \in \mathbb{R}; \quad (3.6)$$

$$E[v(k)x^T(0)] = 0, \quad k \in \mathbb{N}. \quad (3.7)$$

Equation (2.9) implies that $\omega(k)$ is a stationary zero-mean white Gaussian sequence with covariance kernel

$$E[\omega(k)\omega^T(i)] = R_w(T) \delta(k-i), \quad (3.8)$$

$$R_w(T) = \int_0^T \exp(A(T-\tau))R_w \exp(A^T(T-\tau)) d\tau \quad (3.9)$$

The performance index has the form

$$J = \lim_{T \rightarrow \infty} E \left\{ \frac{1}{T} \int_0^T (x^T Q x + u^T R u) d\tau \right\}, \quad (3.10)$$

where $Q \in \mathbb{R}^{n \times n}$, $Q = Q^T$, $Q \geq 0$, $R \in \mathbb{R}^{m \times m}$, $R = R^T$, $R > 0$. The criterion (3.10) can also be computed as follows (see Juan, 1988).

Proposition 1 Suppose a hold function $F(t)$, $t \in [0, T]$ stabilizes asymptotically (2.16). Then the criterion (3.10) has the form

$$J = \lim_{T \rightarrow \infty} E \left\{ \frac{1}{T} \int_{kT}^{(k+1)T} (x^T Q x + u^T R u) d\tau \right\}, \quad (3.11)$$

Proposition 2. Suppose a hold function $F(t)$, $t \in [0, T]$ stabilizes (2.16) asymptotically. Then the criterion (3.10) can be computed as follows

$$J = \frac{1}{T} \text{tr} \{ L_a M_a(T) + R_a N_a(T) + Q P_a(T) \} \quad (3.12)$$

where $L_a \in \mathbb{R}^{(n+q) \times (n+q)}$, $N_a(t) \in \mathbb{R}^{p \times p}$, $P_a(t) \in \mathbb{R}^{n \times n}$ are obtained as follows

$$\dot{D}(t) = AD(t) + BF(t), \quad D(0) = 0 \quad (3.13)$$

$$\begin{aligned} \dot{M}_a(t) = & \begin{bmatrix} \exp(A^T t) + C^T V^T D^T(t) \\ S^T D^T(t) \end{bmatrix} \\ & \times Q \left[\exp(At) + D(t)VC, D(t)S \right] \\ & + \begin{bmatrix} C^T V^T F^T(t) \\ S^T F^T(t) \end{bmatrix} R [F(t)VC, F(t)S]; \quad M_a(0) = 0 \end{aligned} \quad (3.14)$$

$$\begin{aligned} & \begin{bmatrix} \exp(AT) + D(T)VC & D(T)S \\ UC & P \end{bmatrix} \\ & \times L_a \begin{bmatrix} \exp(A^T T) + C^T V^T D^T(T) & C^T U^T \\ S^T D^T(T) & P^T \end{bmatrix} \\ & + \begin{bmatrix} D(T)V \\ U \end{bmatrix} R_v [V^T D^T(T), U^T] + \begin{bmatrix} R_w^{(1)} & 0 \\ 0 & 0 \end{bmatrix} \cdot L_a = 0 \end{aligned} \quad (3.15)$$

$$N_a(t) = \int_0^t [N^T D_w^{(1)}(r) Q D(r) V + V^T F^T(r) R F(r) V] dr \quad (3.16)$$

$$P_a(t) = \int_0^t R_w(r) dr \quad (3.17)$$

Proof. First rewrite (2.10) as (3.13). Define $L_a = \lim_{k \rightarrow \infty} E[x_a(k)x_a^T(k)]$. The stability of (2.16) implies that L_a can be computed by (3.15). Moreover, simple algebraic manipulations based on (2.7)–(2.11) imply that the cost (3.11) can be computed as (3.12) subject to (3.13)–(3.17).

(a) *Free monodromy.*

Proposition 3. A hold function $F(t)$, $t \in [0, T]$ which stabilizes (2.16) asymptotically and minimizes (3.10) must satisfy

$$\begin{aligned} B^T \Psi(t) - 2RF(t)[VC, S]L_a & \begin{bmatrix} C^T & V^T \\ S^T & \end{bmatrix} \\ & - 2RF(t)VR_v V^T = 0 \end{aligned} \quad (3.18)$$

where $\Psi(t) \in \mathbb{R}^{n \times n}$ satisfies

$$\begin{aligned} \dot{\Psi}(t) = & -A^T \Psi(t) + 2Q[\exp(At) + D(t)VC, D(t)S]L_a \\ & \times \begin{bmatrix} C^T & V^T \\ S^T & \end{bmatrix} + 2QD(t)VR_v V^T \end{aligned} \quad (3.19)$$

$$\begin{aligned} \Psi(T) = & -2[L_a, 0]K_a \begin{bmatrix} \exp(AT) + D(T)VC & D(T)S \\ UC & P \end{bmatrix} \\ & \times L_a \begin{bmatrix} C^T & V^T \\ S^T & \end{bmatrix} + \begin{bmatrix} D(T)V \\ U \end{bmatrix} R_v [V^T D^T(T), U^T] \end{aligned} \quad (3.20)$$

$$\begin{aligned} & \begin{bmatrix} \exp(A^T T) + C^T V^T D^T(T) & C^T U^T \\ S^T D^T(T) & P^T \end{bmatrix} \\ & \times K_a \begin{bmatrix} \exp(AT) + D(T)VC & D(T)S \\ UC & P \end{bmatrix} \\ & + M_a(T) - K_a = 0 \end{aligned} \quad (3.21)$$

and the matrices $D(t)$, L_a and $M_a(t)$ are given by (3.13)–(3.15).

Proof. The optimization of (3.12) subject to (3.13)–(3.17) is performed using standard optimal control theory, but where the state variable is the matrix $D(t)$ and the control input is the hold function $F(t)$, yielding (3.18)–(3.21). See Juan (1988) for details.

Equations (3.13)–(3.21) define a two point boundary value problem for the two $n \times r$ matrices $D(t)$ and $\Psi(t)$. The first-order gradient algorithm of Bryson and Ho (1975) has been used successfully to compute solutions to these equations. The examples we have treated suggest that in general, (3.13)–(3.21) always have a solution when the triple (A, B, C) of (2.1)–(2.2) is minimal, but that this solution is not in general unique. However, in the case of fixed monodromy, we can guarantee both existence and uniqueness.

(b) *Fixed monodromy.*

Proposition 4. If the triple (A, B, C) of (2.1)–(2.2) is minimal, and the matrix G of (2.17) is such that the system (2.16) is asymptotically stable, then for almost all T , the hold function $F(t)$, $t \in [0, T]$ which minimizes (3.10) subject to (2.1)–(2.6), (3.1)–(3.9), (2.17) exists and is unique. It satisfies

$$\begin{aligned} D(t) = & AD(t) + \{BR^{-1}B^T\Psi(t) \\ & + \{[VC, S]L_a \begin{bmatrix} C^T & V^T \\ S^T & \end{bmatrix} + VR_v V^T \} \end{aligned} \quad (3.22)$$

$$\begin{aligned} \dot{\Psi}(t) = & -A^T \Psi(t) + 2Q[\exp(At) + D(t)VC, D(t)S]L_a \\ & \begin{bmatrix} C^T & V^T \\ S^T & \end{bmatrix} + 2QD(t)VR_v V^T \end{aligned} \quad (3.23)$$

$$F(t) = \{R^{-1}B^T\Psi(t)\} \begin{bmatrix} [VC, S]L_a \\ \end{bmatrix} + VR_v V^T \quad (3.24)$$

where L_a is the positive semidefinite solution of (3.15).

Proof. See Juan (1988).

Notice that (3.22), (3.23) can be rewritten as a standard Hamiltonian two point boundary value problem for the vectors $\text{vec}(D)$ and $\text{vec}(\Psi)$. The direct solution of (3.22), (3.23) therefore requires only solving linear equations (Bryson and Ho, 1975).

4. *Linear quadratic regulation*

Throughout this section we assume there is no disturbance and no measurement noise: we are regulating the transient behavior of the closed loop system against nonzero initial conditions. In (2.1)–(2.3) we assume

$$w(t) = 0, \quad v(k) = 0, \quad (4.1)$$

$$F(x(0)) = 0, \quad F(\xi(0)) = 0, \quad (4.2)$$

$$F \left(\begin{bmatrix} x(0) \\ \xi(0) \end{bmatrix} \begin{bmatrix} x^T(0) & \xi^T(0) \end{bmatrix} \right) = X_0^* \quad (4.3)$$

and the performance index has the form

$$J = E \left(\int_0^\infty (x^T Q x + u^T R u) dt \right), \quad (4.4)$$

where $Q \in \mathbb{R}^{n \times n}$, $Q = Q^T$, $Q \geq 0$, $R \in \mathbb{R}^{m \times m}$, $R = R^T$, $R > 0$.

(a) *Free monodromy.*

Proposition 5. A hold function $F(t)$, $t \in [0, T]$ which stabilizes asymptotically (2.16) and minimizes (4.4) subject to (2.1)–(2.6), (4.1)–(4.3) must satisfy (3.13)–(3.21) where

$$\begin{aligned} R_w(T) &= 0 \\ 0 &= 0 \end{aligned}$$

is replaced by X_0^* and R_v is replaced by 0. The optimal value of (4.4) is then

$$J = \text{tr}(K_a X_0^*) \quad (4.5)$$

(b) *Fixed monodromy.*

Proposition 6. If the triple (A, B, C) of (2.1)–(2.2) is minimal and the matrix G of (2.17) is such that the system

(2.16) is asymptotically stable then for almost all T , the hold function $F(t)$, $t \in [0, T]$ which minimizes (4.4) subject to (2.1)–(2.6), (4.1)–(4.3), (2.17) exists and is unique. It satisfies (3.22)–(3.24) where

$$R_{\infty}(T) = 0$$

is replaced by X_0^* and R_{∞} by 0. The optimal value of (4.4) is then given by (4.5).

Remark 1. Propositions 5 and 6 reveal that the linear quadratic regulator is a particular case of linear quadratic Gaussian regulator where $R_{\infty} = 0$ and X_0^* replaces

$$R_{\infty}(T) = 0$$

In other words, optimizing the transient performance (4.4) of a noiseless undisturbed system is equivalent to optimizing the steady state performance (3.11) of the same noiseless system under some properly defined perturbations.

Remark 2. Propositions 3, 4, 5 and 6 illustrate the fact that fixed monodromy problems are easier to solve than free monodromy problems. Not only can we guarantee existence and uniqueness of an optimal fixed monodromy hold function, but it can be computed directly at the expense of solving linear equations. On the other hand, for free monodromy problems we cannot guarantee existence of a solution, neither can we guarantee its uniqueness, nor can we guarantee that the iterative algorithm we used will always converge. The hierarchy of difficulty between free and fixed monodromy problems is reminiscent of the problem of optimal L^2 model reduction (Wilson, 1970) where, if the poles of the optimal reduced order model are free, it is not guaranteed to exist, nor to be unique, nor to be computable by a convergent algorithm; whereas if these poles are fixed, the reduced order model is computed directly by solving linear equations

5. Example

In this section we illustrate the use of the results of Sections 3–4. We compare two sampled data regulators for a given plant with respect to an integral quadratic performance index. The first regulator is obtained by using a standard zero-order hold, preceded by a discrete compensator based on pole assignment. The second regulator is obtained by using the same discrete compensator, but optimizing the hold function with respect to the performance index, under the “fixed monodromy constraint” that the discretized plant in the second configuration be the same as the discretized plant in the first configuration. In other words, we show that it is possible to improve the performance of a feedback loop by adjusting the hold function, without changing the discrete compensator nor the discrete model of the plant. Such an improvement is then due exclusively to a better intersam-

pling behavior. Note that the discrete compensator has not been chosen for optimality with respect to the performance index, because this optimization would depend on the hold function, and the effect of optimizing the hold function alone would then be difficult to assess.

The plant is a simple harmonic oscillator with position measurement of the form:

$$\dot{x} = \begin{bmatrix} 0 & -5 \\ 5 & 0 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u$$

The sampling period is $T = 1$.

(A) The first controller is implemented using a dynamic compensator and a zero-order hold

$$u(t) = \gamma(k), \quad t \in [kT, (k+1)T) \tag{5.1}$$

The discretized model of the plant is then

$$x(k+1) = \begin{bmatrix} 0.2837 & 0.9589 \\ -0.9589 & 0.2837 \end{bmatrix} x(k) + \begin{bmatrix} -0.1433 \\ -0.1918 \end{bmatrix} \gamma(k)$$
$$\zeta(k) = [1 \ 0] x(k) \tag{5.3}$$

The dynamic compensator is

$$\xi(k+1) = \begin{bmatrix} -0.06732 & 0.3735 \\ 0.2057 & -0.5 \end{bmatrix} \xi(k) + \begin{bmatrix} 0.5673 \\ 0.875 \end{bmatrix} \zeta(k)$$
$$\gamma(k) = [-1.5094 \ 4.086] \xi(k) \tag{5.4}$$

and has been chosen so that the closed loop eigenvalues are located at $\pm 1.0 \times 10^{-3}$, and $\pm j \times 10^{-3}$. The cost function has the form (4.4), (4.3) where

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$R = 1$$
$$X_0^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.5}$$

For this regulator, the performance index is computed using Proposition 2 which yields

$$J_1 = 47.82 \tag{5.6}$$

(B) The second controller consists of a sampler with generalized hold function, together with the same discrete compensator (5.4). The hold function is constrained such that the discretized plant is the same as (5.3). In other words, the closed loop systems A and B have the same dynamics at sampling instants. However, in case B, the hold function is chosen so as to optimize the cost function (4.4), (4.3), (5.5). Applying Proposition 6 yields the hold function of Fig. 2, with

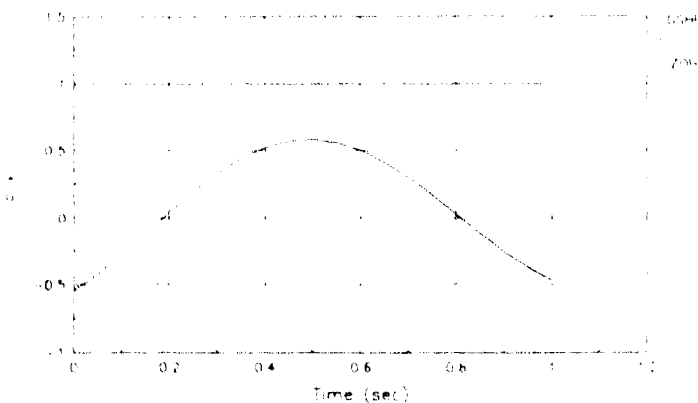


FIG. 2. Generalized sampled data hold functions

cost function

$$J_2 = 25.43 \quad (5.7)$$

which represents a 47% improvement over (5.6)

6 Conclusions

The main contribution of this paper has been to present methods for optimizing the intersampling behavior of sampled data control systems. The basic premise is that the hold function itself is the design variable. We have presented explicit solutions to the optimal LQ and LQG regulation problems. It is found that for fixed monodromy problems, we can not only guarantee existence and uniqueness of an optimal hold function, but we can compute it directly by solving a linear two point boundary value problem of Hamiltonian structure. This brief paper has only presented the basic theoretical results on optimal sampled data regulation by GSHF. Engineering design issues and detailed examples are considered in Yang and Kabamba (1988) where it is also shown that GSHF control can be used to achieve simultaneous quadratic performance in several control loops using the same controller.

References

- Berger, C. S. (1985). The use of intersampling in adaptive control. *IEEE Trans. Aut. Control*, **AC-30**.
- Bryson, A. E. and Y.-C. Ho (1975). *Applied Optimal Control*. Hemisphere, Washington.
- Chammas, A. B. and C. T. Leondes (1978a). On the design of linear time invariant systems by periodic output feedback, Parts I and II. *Int. J. Control*, **27**, 885-903.
- Chammas, A. B. and C. T. Leondes (1978b). Optimal control of stochastic linear systems by discrete output feedback. *IEEE Trans. Aut. Control*, **AC-23**, 921-926.
- Chammas, A. B. and C. T. Leondes (1979). On the finite time control of linear systems by piecewise constant output feedback. *Int. J. Control*, **30**, 227-234.
- De Souza, C. E. and G. C. Goodwin (1984). Intersample variances in discrete minimum variance control. *IEEE Trans. Aut. Control*, **AC-29**.
- Dorato, P. and A. H. Levis (1971). Optimal linear regulators: The discrete-time case. *IEEE Trans. Aut. Control*, **AC-16**, 613-620.
- Franklin, G. F. and J. D. Powell (1980). *Digital Control of Dynamic Systems*. Addison-Wesley, Reading, MA.
- Franklin, G. F. and A. Emami-Naeini (1986). Design of ripple-free multivariable robust servomechanisms. *IEEE Trans. Aut. Control*, **AC-31**.
- Greshak, J. P. and G. C. Verghese (1982). Periodically varying compensation of time-invariant systems. *Syst. Control Lett.*, **2**, 88-93.
- Juan, Y.-C. and P. T. Kabamba (1987). Constrained structure control of periodic systems using generalized sampled-data hold functions. *Proc. 1987 ACC*, pp. 1695-1700, Minneapolis, MN.
- Juan, Y.-C. (1988). Control of linear systems with generalized hold functions. Ph.D. Dissertation, The University of Michigan, MI.
- Kabamba, P. T. (1986). Monodromy eigenvalue assignment in linear periodic systems. *IEEE Trans. Aut. Control*, **AC-31**, 950-952.
- Kabamba, P. T. (1987). Control of linear systems using generalized sampled-data hold functions. *IEEE Trans. Aut. Control*, **AC-32**, 772-783.
- Kaczorek, T. (1985). Pole assignment for linear discrete-time systems by periodic output feedbacks. *Syst. Control Lett.*, **6**, 267-269.
- Kaczorek, T. (1986). Deadbeat control of linear discrete-time systems by periodic output feedback. *IEEE Trans. Aut. Control*, **AC-31**, 1153-1156.
- Khargonekar, P. P., K. Poolla and H. Tannenbaum (1985). Robust control of linear time-invariant plants using periodic compensation. *IEEE Trans. Aut. Control*, **AC-30**, 1088-1096.
- Sirisen, H. R. (1985). Ripple-free deadbeat control of SISO discrete systems. *IEEE Trans. Aut. Control*, **AC-30**.
- Tarn, T. J., J. R. Zavren and X. Zeng (1988). Stabilization of infinite dimensional systems with periodic feedback gains and sampled output. *Automatica*, **24**, 95-99.
- Ukura, S. and A. Nagata (1987). Ripple free deadbeat control for sampled-data systems. *IEEE Trans. Aut. Control*, **AC-32**.
- Wilson, D. A. (1970). Optimal solution of model reduction problems. *Proc. IEEE*, **17**, 1161-1165.
- Yang, C. and P. T. Kabamba (1988). Simultaneous design of sampled data control systems. In *Proc. 1988 ACC*, Atlanta, GA, to appear in *IEEE Trans. Aut. Control*.
- Zavren, J. R. (1983). Stabilisation of infinite-dimensional systems via periodic output feedback. Sc.D. Dissertation, Washington University, Saint Louis, MO.
- Zavren, J. R. and T. J. Tarn (1984). Periodic output feedback stabilization of infinite dimensional systems. *Proc. 9th IFAC Triennial World Congress*, pp. 1445-1449, Budapest, Hungary.
- Zang, X. (1985). Sampled output stabilization of infinite-dimensional systems. Sc.D. Dissertation, Washington University, Saint Louis, MO.

Brief Paper

An Indirect Prediction Error Method for System Identification*

T. SÖDERSTRÖM,[†] P. STOICA[‡] and B. FRIEDLANDER[§]

Key Words—System identification; parameter estimation; prediction error method; model structure; maximum likelihood methods; generalized least squares

Abstract—A new form of prediction error method (PEM) is developed. It is applicable to the case where the model structure of interest can be imbedded in a larger model structure whose estimation is relatively easy. An optimal way of reducing the larger model to the smaller model structure is presented and various interpretations of this reduction are given. The proposed method will have the same asymptotic statistical properties as the standard PEM but it can be implemented by a more efficient algorithm. The properties of the method are illustrated by the means of some simulated examples.

1. Introduction

PREDICTION ERROR methods (PEMs) (Ljung, 1976), are now well-known tools to get statistically efficient parametric models in system identification. Considerable interest has been devoted to developing alternative methods and algorithms which require a lesser amount of computation without sacrificing too much of the statistical efficiency. This paper follows this line of research.

Consider the case of two nested model structures M_1 and M_2 , i.e. let M_1 be a subset of M_2 . (Illustrations are given in Examples 1.1 and 1.2 below.) It is assumed that a PEM is relatively easy to apply in M_2 . We will present and analyze an identification method, which produces a model in M_1 using the PEM model parameters in M_2 as the "data" for this estimation. Due to the two step procedure

measured data $\xrightarrow{\text{(PEM)}} \text{model in } M_2 \rightarrow \text{model in } M_1$

we will call the proposed method an *indirect prediction error method* (IPEM). The IPEM estimates will be shown to have similar statistical properties to the PEM estimates in M_1 . In some circumstances they can be computed much more efficiently.

Note that the second step in the above scheme makes sense, even though M_2 itself may provide a reasonable model for the system under study. According to the parsimony principle [see for example Söderström and Stoica (1989) for a general treatment and Stoica and Söderström (1982), Stoica *et al.* (1985a) for some specialized analysis], some accuracy is lost if an unnecessarily complex model structure is used (such as M_2 instead of M_1).

We assume that both model structures, M_1 and M_2 , are given *a priori*. This might not be the case in practice. For determining an appropriate model structure M_1 one can apply the approach proposed in this paper to models of various orders, and validate them with standard means, see, for example, Söderström and Stoica (1989). The "best" model so obtained can then be taken as the final one. Similarly, an appropriate model order for M_2 can be determined by standard methods.

Next let us exhibit two typical cases of nested or hierarchical model structures M_1 and M_2 . In both cases we will let θ denote the parameter vector in M_1 while M_2 is parametrized by a vector α .

Example 1.1 ("The generalized least squares structure") Let M_1 be given by

$$M_1: A_1(q^{-1})y(t) = B_1(q^{-1})u(t) + \frac{1}{C_1(q^{-1})}e(t) \quad (1.1a)$$

In (1.1) $u(t)$ denotes the input signal, $y(t)$ the output signal and $e(t)$ white noise. The polynomials and the parameter vector θ are

$$\begin{aligned} A_1(q^{-1}) &= 1 + a_1^{(1)}q^{-1} + \dots + a_{n_a}^{(1)}q^{-n_a} \\ B_1(q^{-1}) &= b_1^{(1)}q^{-1} + \dots + b_{n_b}^{(1)}q^{-n_b} \\ C_1(q^{-1}) &= 1 + c_1^{(1)}q^{-1} + \dots + c_{n_c}^{(1)}q^{-n_c} \end{aligned} \quad (1.1b)$$

$$\theta = [a_1^{(1)} \dots a_{n_a}^{(1)} \ b_1^{(1)} \dots b_{n_b}^{(1)} \ c_1^{(1)} \dots c_{n_c}^{(1)}]^T \quad (1.1c)$$

and q^{-1} denotes the backward shift operator. A model of this form is used for example in the generalized least squares (GLS) method, cf. Clarke (1967), Söderström (1974).

By multiplying (1.1a) with $C_1(q^{-1})$ we can rewrite M_1 as

$$M_1: A_1(q^{-1})C_1(q^{-1})y(t) = B_1(q^{-1})C_1(q^{-1})u(t) + e(t) \quad (1.1d)$$

This model is a linear regression *with constraints*, since the polynomial operators have $C_1(q^{-1})$ as a common factor. By neglecting this constraint we arrive at the model structure M_2

$$M_2: A_2(q^{-1})y(t) = B_2(q^{-1})u(t) + e(t) \quad (1.1e)$$

where $\deg A_2 = n_a + n_c$, $\deg B_2 = n_b + n_c$. The parameter vector α is

$$\alpha = [a_1^{(2)} \dots a_{n_a}^{(2)} \ b_1^{(2)} \dots b_{n_b}^{(2)}] \quad (1.1f)$$

Example 1.2 ("AR process observed in noise") Let M_1 model an autoregressive process observed in additive white noise. Such a model is frequently used in various signal processing applications. It can be written as

$$M_1: y(t) = \frac{1}{A(q^{-1})}v(t) + w(t) \quad (1.2a)$$

Let the white noise sequences $v(t)$ and $w(t)$ have variances σ_v^2 and σ_w^2 , respectively. The parameter vector θ is given by

$$\theta = [a_1 \dots a_n \ \sigma_v^2 \ \sigma_w^2] \quad (1.2b)$$

* Received 14 October 1988; revised 24 April 1989; received in final form 25 April 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor D. W. Clarke under the direction of Editor P. C. Parks.

[†] Automatic Control and Systems Analysis Group, Department of Technology, Uppsala University, PO Box 534, S-751 21 Uppsala, Sweden. Author to whom all correspondence should be addressed.

[‡] Faculty of Automation, Bucharest Polytechnic Institute, Splaiul Independenței 313, R-77206 Bucharest, Romania.

[§] Signal Processing Technology, 703 Coastland Drive, Palo Alto, CA 94303, U.S.A.

The spectral density of $y(t)$ is

$$\phi_y(\omega) = \frac{\sigma_v^2 + \sigma_w^2 A(e^{j\omega})A(e^{-j\omega})}{A(e^{j\omega})A(e^{-j\omega})} \tag{1.2c}$$

By using spectral factorization, the measurements $y(t)$ can also be represented by an ARMA model

$$M_2: A(q^{-1})y(t) = C(q^{-1})e(t) \tag{1.2d}$$

where $C(q^{-1}) = 1 + c_1q^{-1} + \dots + c_nq^{-n}$ and $\sigma_e^2 = Ee^2(t)$ are given by

$$\sigma_e^2 C(e^{j\omega})C(e^{-j\omega}) = \sigma_v^2 + \sigma_w^2 A(e^{j\omega})A(e^{-j\omega}). \tag{1.2e}$$

The parameter vector α for M_2 is

$$\alpha = [a_1 \dots a_n \ c_1 \dots c_n \ \sigma_v^2]^\top \tag{1.2f}$$

Both models, (1.2a) and (1.2d) are valid representations of $y(t)$. However, (1.2d) has the drawback of being a nonparsimonious representation [it has $n + 1$ parameters in addition to (1.2a)]. Furthermore, for most applications the parameters of interest are (1.2b) (in particular the signal parameters a_i and σ_v^2), not the parameters of the ARMA representation of the signal-to-noise process.

Other examples of nested model structures are described by Dasgupta *et al.* (1988). They treat the situation when θ corresponds to a set of physical parameters and show that in many cases α is related to θ in a so called multilinear fashion. Similarly, Bastin *et al.* (1989) give cases where θ describes physical parameters. These authors show that for a large class of models, the model M_1 can be written as a nonlinear regression and imbedded in a linear regression model in M_2 .

The IPEM can be viewed as a way for *model reduction*, since the larger model in M_2 is reduced in Step 2. A similar principle, although not based on a PEM, is suggested by Porat and Friedlander (1986) and Rosen and Porat (1986) for time series analysis. In these references, fitting a model in M_2 consists of estimating the covariance function of the signal, which can be done by a simple procedure. In the second stage a parametric model is fitted to the covariance function data. A similar approach for ARMA modelling is proposed by Stoica *et al.* (1987). Moses (1986) gives a method where the covariance function of a time series is estimated in a first step while techniques for approximate realization and model reduction are used in the second step. Wahlberg (1986, 1987, 1989) has proposed an alternative approach in which M_2 corresponds to an autoregressive model of large order. Such a model can be computed easily since it is the solution of a linear least-squares problem. Furthermore, the problem of reducing identified models has also been discussed by Jakeman and Young (1981, 1983). A situation similar to the previously introduced problem occurs when using the approach of *indirect identification* to estimate the parameters of a system under feedback control see Ljung *et al.* (1974) or Soderström and Stoica (1989) for details. In such a case the closed loop system is first identified (which in the framework of this paper corresponds to fitting a model in M_2 to the data). In a second step the open loop dynamics is solved for, assuming the feedback is known. Since the open loop system typically is of lower order than the closed loop system, we have a model reduction from M_2 to M_1 .

The paper is organized as follows. In the next section the IPEM is introduced formally and some interpretations of its second step, i.e. reduction of M_2 to M_1 , are given. In Section 3 the asymptotic distribution of the IPEM estimates is established, while Section 4 is devoted to implementation issues of the IPEM.

2 The indirect prediction error method

For a detailed discussion of prediction error methods and their role in system identification, see Ljung (1976, 1987), Söderström and Stoica (1989). The PE estimates of θ and α are given by

$$\hat{\theta} = \arg \min_{\theta} \sum_{t=1}^N \epsilon_1^2(t, \theta) \tag{2.1a}$$

$$\hat{\alpha} = \arg \min_{\alpha} \sum_{t=1}^N \epsilon_2^2(t, \alpha) \tag{2.1b}$$

where N is the number of data points, and $\epsilon_1(t, \theta)$ is the prediction error

$$\epsilon_1(t, \theta) = y(t) - \hat{y}(t | t - 1; \theta)$$

in the model structure M_1 and similarly for $\epsilon_2(t, \alpha)$

We will impose the following two assumptions on the model structures M_1 and M_2 and the data.

A1. The structures are nested

$$M_1 \subseteq M_2 \tag{2.2}$$

so that there exists a smooth map

$$\alpha(\theta): S_1 \rightarrow S_2, \quad S_1 \subseteq R^{dim \theta}, \quad S_2 \subseteq R^{dim \alpha}$$

such that

$$\epsilon_2(t, \alpha(\theta)) = \epsilon_1(t, \theta) \tag{2.3}$$

Furthermore, $\dim \theta \leq \dim \alpha$ and S_1 is an open (nonzero measure) set in $R^{dim \theta}$ such that $\epsilon_1(t, \theta)$ is a stationary process for any $\theta \in S_1$. We assume that $\partial \alpha(\theta) / \partial \theta$ has full rank over S_1 .

A2. Both structures give parameter identifiability. This means that there exist *unique*, "true", parameter vectors α^* and θ^* such that

$$\epsilon_1(t, \theta^*) = \epsilon_2(t, \alpha^*) \text{ is white noise of zero mean and variance } \lambda^2, \quad \lambda > 0. \tag{2.4}$$

Note that (2.4) and assumption A1 imply

$$\alpha^* = \alpha(\theta^*) \tag{2.5}$$

Assumption A1 is solely a condition on M_1 and M_2 , while A2 involves also the properties of the data ("the system"). Assumption A2 may seem strong. The requirement that the quantity in (2.4) is white noise is neither needed for derivation of the estimator (2.9) nor the interpretations 2.1 and 2.3 to follow. It is needed though for the ML interpretation 2.2 and in the analysis of asymptotic properties (Section 3).

Example 2.1 [The map $\alpha(\theta)$]. In Example 1.1 the map $\alpha(\theta)$ is in implicit form given by

$$A_2(z) = A_1(z)C_1(z) = B_2(z) = B_1(z)C_1(z) \tag{2.6a}$$

In this case

$$\dim \alpha = na_2 + nb = na_1 + nb_1 + 2nc_1 = na_1 + nb_1 + nc_1 = \dim \theta \tag{2.6b}$$

In Example 1.2 the map $\alpha(\theta)$ is implicitly given by the identity, (1.2c).

$$\sigma_e^2 C(z)C(z^{-1}) = \sigma_v^2 + \sigma_w^2 A(z)A(z^{-1}) \tag{2.6c}$$

Here

$$\dim \alpha = 2n + 1 \geq n + 2 = \dim \theta \tag{2.6d}$$

Since M_1 is a subset of M_2 (and hence a more parsimonious model structure) it will give a more accurate model than M_2 according to the parsimony principle, see for example Soderström and Stoica (1989). Therefore from the *accuracy* standpoint we should be interested in the estimation of $M_1(\theta)$. Now, it may happen that $\hat{\theta}$, (2.1a), is more difficult to compute than $\hat{\alpha}$, (2.1b). This is clearly the case in Example 1.1 since M_2 is a *linear* regression.

Due to the above considerations it may be preferable to first obtain the estimate $\hat{\alpha}$ of α and then use this estimate to determine an estimate, say $\hat{\theta}$, of θ .

If S_2 is an open (nonzero measure) set in $R^{dim \alpha}$ we can expect that $\hat{\alpha} \in S_2$ (at least for large N), since $\alpha^* \in S_2$. In such a situation we can compute $\hat{\theta}$ as

$$\hat{\theta} = \theta(\hat{\alpha}) \tag{2.7}$$

where $\theta(\alpha)$ is the inverse function

$$\theta(\alpha): S_2 \rightarrow S_1$$

It is easy to see that $\hat{\theta} = \hat{\theta}$. Indeed

$$\sum_{i=1}^N \epsilon_i^2(t, \theta) = \sum_{i=1}^N \epsilon_i^2(t, \alpha(\theta)) \geq \sum_{i=1}^N \epsilon_i^2(t, \hat{\alpha}) = \sum_{i=1}^N \epsilon_i^2(t, \hat{\theta}) \quad \text{for any } \theta \in S_1 \quad (2.8)$$

which implies $\hat{\theta} = \hat{\theta}$. We used the assumption that $\hat{\alpha} \in S_2$ in the last equality in (2.8).

However, in most cases of interest $\dim \theta < \dim \alpha$ and S_2 is a "thin" set of $R^{\dim \alpha}$. Then $\hat{\alpha} \notin S_2$ (w.p.1) and (2.7) cannot be used. In such a case we could determine θ as the solution to the following optimization problem:

$$\hat{\theta} = \arg \min_{\theta} V(\theta) \quad V(\theta) = \frac{1}{2} [\hat{\alpha} - \alpha(\theta)]^T \hat{P}_\alpha^{-1} [\hat{\alpha} - \alpha(\theta)]$$

where \hat{P}_α is a consistent estimate of

$$P_\alpha = E \left(\frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha} \frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha}^T \right) \quad (2.9)$$

A natural estimate is

$$\hat{P}_\alpha = \frac{1}{N} \sum_{i=1}^N \left(\frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha} \frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha}^T \right) \quad (2.10)$$

Hence the indirect prediction error method (IPEM) consists of the following two steps:

Step 1. Determine $\hat{\alpha}$ from (2.1b) and \hat{P}_α .

Step 2. Determine θ from (2.9).

Before giving some motivations of the IPEM recall that under A1-A2, the normalized estimation errors $\sqrt{N}(\hat{\theta} - \theta^*)/\lambda$ and $\sqrt{N}(\hat{\alpha} - \alpha^*)/\lambda$ are asymptotically normally distributed with zero means and covariance matrices given by

$$P_\theta = E \left(\frac{\partial \epsilon_i(t, \theta)}{\partial \theta} \frac{\partial \epsilon_i(t, \theta)}{\partial \theta}^T \right) \quad (2.11)$$

and (2.10), respectively. (Ljung, 1987; Soderstrom and Stoica, 1989). Furthermore, it is not difficult to see that

$$P_\alpha = [G P_\theta^{-1} G^T]^{-1} \quad (2.12a)$$

where

$$G = \frac{\partial \alpha(\theta)}{\partial \theta} \quad (2.12b)$$

and where we have assumed G to have full rank.

2.1 *A geometrical interpretation* This idea is given in Soderstrom (1975) and Stoica (1976) (and see also Dasgupta, 1988), and can be described as follows. The optimization problem (2.9) can be alternatively stated as

$$\hat{\alpha} = \arg \min_{\alpha \in S_2} (\hat{\alpha} - \alpha)^T \hat{P}_\alpha^{-1} (\hat{\alpha} - \alpha) \quad (2.13a)$$

$$\hat{\theta} = \theta(\hat{\alpha}) \quad (2.13b)$$

The geometrical interpretation then becomes quite clear. The idea is to determine the point α in S_2 which is closest (in the metric induced by \hat{P}_α) to $\hat{\alpha}$, and to use $\hat{\alpha}$ and the inverse function $\theta(\alpha)$ as in (2.7) to obtain θ .

2.2 *A maximum likelihood interpretation* This idea is based on the development in Stoica *et al.* (1985b). Consider $\hat{\alpha}$ computed in Step 1 as a "statistic". The asymptotic log-likelihood function of $\hat{\alpha}$ is given by

$$L(\theta) = -\frac{\dim \alpha}{2} \log(2\pi) - \frac{1}{2} \log \det(\lambda^2 P_\alpha/N)$$

$$- \frac{N}{2\lambda^2} [\hat{\alpha} - \alpha(\theta)]^T P_\alpha^{-1} [\hat{\alpha} - \alpha(\theta)] \quad (2.14a)$$

The derivative of $L(\theta)$ with respect to θ satisfies

$$\frac{\partial^2 L(\theta)}{\partial \theta^2} = -\frac{\lambda^2}{2N} \frac{\partial}{\partial \theta} \log \det P_\alpha - \frac{\partial}{\partial \theta} \left[\frac{1}{2} \|\hat{\alpha} - \alpha(\theta)\|_{P_\alpha}^2 \right]$$

$$= -\frac{1}{2} [\hat{\alpha} - \alpha(\theta)]^T \frac{\partial P_\alpha}{\partial \theta} [\hat{\alpha} - \alpha(\theta)] \quad (2.14b)$$

where the last term is written in an informal way. The second term in (2.14b) is related to $-(\partial/\partial \theta)V(\theta)$, with $V(\theta)$ given by (2.9). Provided L has a maximum at θ close to θ^* (which is true for N large) then it follows from (2.14b) that

$$\frac{\partial}{\partial \theta} V(\theta)|_{\theta=\theta^*} = O(1/N) \approx 0 \quad (2.15a)$$

Hence the estimate θ given by (2.9) is a good large sample approximation of θ in the sense

$$\theta - \theta^* = O(1/N) \quad (2.15b)$$

A related approach has been proposed by Erdlander and Porat (1988). It concerns estimating parameters in a model based on a given statistic. In terms of the problem in this paper the formulation is as follows. Let $\hat{\alpha}$ be a consistent estimate of α with a covariance matrix $Q(\theta)$. Consider estimates of the form

$$\hat{\theta} = f(\hat{\alpha}) \quad (2.16a)$$

where the function $f(\alpha)$ is left inverse to $\alpha(\theta)$, see (2.5), in the sense

$$\theta = f(\alpha(\theta)) \quad (2.16b)$$

Note that since $\dim \alpha = \dim \theta$ there are many such functions f . Then the covariance matrix of $\hat{\theta}$ has, for N large, a lower bound. This bound is achieved if the estimate $\hat{\theta}$ (2.16a) is chosen as

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{2} [\hat{\alpha} - \alpha(\theta)]^T Q^{-1}(\theta) [\hat{\alpha} - \alpha(\theta)] \quad (2.16c)$$

If $Q(\theta)$ in (2.16c) is replaced by the estimate \hat{P}_α (such a replacement should have only a higher order effect on $\hat{\theta}$) we arrive at the estimate $\hat{\theta}$ (2.9).

2.3 *A loss function interpretation* Let θ be close to θ^* [given by (2.1a)]. Since the function $\alpha(\theta)$ is smooth $\alpha(\theta)$ will then be close to α . A Taylor series expansion of $\epsilon(t, \alpha(\theta))$ gives

$$\epsilon(t, \alpha(\theta)) = \epsilon(t, \alpha) + \frac{\partial \epsilon(t, \alpha)}{\partial \alpha} (\alpha(\theta) - \alpha) \quad (2.17a)$$

Since, due to

$$0 = \frac{1}{N} \sum_{i=1}^N \epsilon_i(t, \alpha) \quad \frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha} = 0 \quad (2.17b)$$

we get

$$\frac{1}{2N} \sum_{i=1}^N \epsilon_i^2(t, \alpha(\theta)) = \frac{1}{2N} \sum_{i=1}^N \epsilon_i^2(t, \alpha) + \frac{1}{2} [\alpha - \alpha(\theta)]^T$$

$$\left[\frac{1}{N} \sum_{i=1}^N \left(\frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha} \frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha}^T \right) \right] [\alpha - \alpha(\theta)]$$

$$+ \frac{1}{2} \left[\frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha} \frac{\partial \epsilon_i(t, \alpha)}{\partial \alpha}^T \right]_{\alpha=\alpha(\theta)} [\alpha - \alpha(\theta)]$$

$$= \text{constant} + V(\theta) \quad (2.17c)$$

Hence for θ close to θ^* the loss function in (2.1a) behaves approximately as $V(\theta)$. In fact, (2.17c) will hold exactly (regardless the distance of θ from θ^*) if

(i) $\epsilon(t, \alpha)$ is affine in α . This implies that (2.17a) is exact.

(ii) The estimate \hat{P}_α given in (2.11), is used.

In such a case

$$\hat{\theta} = \theta^* \quad (2.17d)$$

for any finite N .

Note that for the GLS case (Example 1.1), the two conditions (i) and (ii) above are satisfied. The relation (2.17d) was established for this case in Soderstrom (1977). See also Stoica and Soderstrom (1989) for a further discussion of this interpretation.

3. Asymptotic properties of the IPEM

In (2.17d) we established that under certain conditions the IPEM estimate $\hat{\theta}$ coincides exactly for finite N with the PEM estimate θ . This is a very strong property that holds only in

special situations. However, under much weaker assumptions the estimates $\hat{\theta}$ and $\bar{\theta}$ have the same asymptotic distribution as will now be shown. Note that this is an attractive property since the estimates then will have the same statistical properties.

Let N be large. Then we have

$$\begin{aligned}\hat{\alpha} - \alpha^* &\approx \hat{\alpha} - \alpha(\theta^*) = O(1/\sqrt{N}) \\ V(\theta) &\approx V(\theta^*) = O(1/N) \\ \hat{\alpha} - \alpha(\hat{\theta}) &= O(1/\sqrt{N}) \\ \alpha(\hat{\theta}) - \alpha(\theta^*) &= O(1/\sqrt{N}) \\ \hat{\theta} - \theta^* &= O(1/\sqrt{N})\end{aligned}$$

Hence the estimate θ is root- N consistent. By series expansion we get

$$\begin{aligned}0 &= V_{\alpha}^1(\theta) = V_{\alpha}^1(\theta^*) + V_{\alpha\alpha}(\theta^*)(\hat{\theta} - \theta^*) + O(1/N) \\ \hat{\theta} - \theta^* &= -V_{\alpha\alpha}^{-1}(\theta^*)V_{\alpha}^1(\theta^*) + O(1/N)\end{aligned}\tag{3.1}$$

Differentiating $V(\theta)$ gives

$$\begin{aligned}V_{\alpha}(\theta) &\approx -[\hat{\alpha} - \alpha(\theta)]^T \hat{P}_{\alpha}^{-1} \frac{\partial \alpha(\theta)}{\partial \theta} \\ V_{\alpha\alpha}(\theta) &\approx \left[\frac{\partial \alpha(\theta)}{\partial \theta} \right]^T \hat{P}_{\alpha}^{-1} \frac{\partial \alpha(\theta)}{\partial \theta} - [\hat{\alpha} - \alpha(\theta)]^T \hat{P}_{\alpha}^{-1} \frac{\partial^2 \alpha}{\partial \theta^2}\end{aligned}\tag{3.2}$$

The term $(\partial^2 \alpha / \partial \theta^2)$ is written in an informal way. Strictly speaking it is tensor. The second term of $V_{\alpha\alpha}(\theta)$, when evaluated at $\theta = \theta^*$, can be neglected. From (3.2) we obtain

$$\begin{aligned}V_{\alpha}(\theta^*) &\approx -(\hat{\alpha} - \alpha^*)^T \hat{P}_{\alpha}^{-1} G^T = -(\hat{\alpha} - \alpha^*)^T P_{\alpha}^{-1} G^T + O(1/N) \\ V_{\alpha\alpha}(\theta^*) &\approx G \hat{P}_{\alpha}^{-1} G^T + O(1/\sqrt{N}) = G P_{\alpha}^{-1} G^T + O(1/\sqrt{N})\end{aligned}\tag{3.3}$$

Now (3.1) and (3.3) imply

$$\hat{\theta} - \theta^* = (G P_{\alpha}^{-1} G^T)^{-1} G P_{\alpha}^{-1} (\hat{\alpha} - \alpha^*) + O(1/N)\tag{3.4}$$

and hence

$$\frac{\sqrt{N}}{\lambda} (\hat{\theta} - \theta^*) \xrightarrow{\text{dist}} \mathbf{N}(0, P)\tag{3.5a}$$

with

$$\begin{aligned}P &= (G P_{\alpha}^{-1} G^T)^{-1} G P_{\alpha}^{-1} \lim_{N \rightarrow \infty} \text{cov} \left(\frac{\sqrt{N}}{\lambda} (\hat{\alpha} - \alpha^*) \right) \\ &\quad \times P_{\alpha}^{-1} G^T (G P_{\alpha}^{-1} G^T)^{-1} \\ &= (G P_{\alpha}^{-1} G^T)^{-1} = P_{\theta}\end{aligned}\tag{3.5b}$$

cf (2.13a). This means that $\hat{\theta}$ and $\bar{\theta}$ have the same asymptotic distribution.

4. Implementation of the IPEM

In this section we will first discuss how to perform Step 2 of the IPEM, i.e. the minimization of the loss function $V(\theta)$, (2.9). We develop a Gauss-Newton (GN) algorithm for this purpose. For a sufficiently large amount of data, this algorithm will converge in one step provided it is initialized by a consistent estimate $\theta^{(0)}$ of θ^* . Let $\hat{\theta}^{(k)}$ denote the estimate at iteration k . By dropping the second term of $V_{\alpha\alpha}(\theta)$ in (3.2) we get the following algorithm from a standard Newton method:

$$\begin{aligned}\hat{\theta}^{(k+1)} &= \hat{\theta}^{(k)} + \{G(\hat{\theta}^{(k)}) \hat{P}_{\alpha}^{-1} G^T(\hat{\theta}^{(k)})\}^{-1} \\ &\quad \times \{G(\hat{\theta}^{(k)}) \hat{P}_{\alpha}^{-1} [\hat{\alpha} - \alpha(\hat{\theta}^{(k)})]\}\end{aligned}\tag{4.1}$$

If $\hat{\theta}^{(0)}$ is a consistent estimate of θ^* we can assume that

$$|\hat{\theta}^{(0)} - \theta^*| = O(1/\sqrt{N}).\tag{4.2}$$

We then get

$$\begin{aligned}\hat{\theta}^{(1)} - \theta^* &= \hat{\theta}^{(0)} - \theta^* + \{G(\hat{\theta}^{(0)}) \hat{P}_{\alpha}^{-1} G^T(\hat{\theta}^{(0)})\}^{-1} \\ &\quad \times \left\{ G(\hat{\theta}^{(0)}) \hat{P}_{\alpha}^{-1} \left[\hat{\alpha} - \alpha(\theta^*) \right. \right. \\ &\quad \left. \left. + G^T(\hat{\theta}^{(0)} - \theta^*) + o\left(\frac{1}{N}\right) \right] \right\}\end{aligned}$$

$$\begin{aligned}&= (G P_{\alpha}^{-1} G^T)^{-1} G P_{\alpha}^{-1} (\hat{\alpha} - \alpha^*) + o\left(\frac{1}{N}\right) \\ &= \bar{\theta} - \theta^* + o\left(\frac{1}{N}\right).\end{aligned}\tag{4.3}$$

In the last line we have used (3.4). The result (4.3) means that $\hat{\theta}^{(1)}$ and $\bar{\theta}$ are asymptotically equal and that the GN algorithm (4.1) for large N converges in one iteration only.

Let us now discuss the choice of the initial value $\hat{\theta}^{(0)}$ for the optimization. A consistent estimate $\hat{\theta}^{(0)}$ can be constructed in many ways. Note that we would like to obtain $\hat{\theta}^{(0)}$ without reprocessing the observed samples $y(1) \cdots y(N)$. Indeed the computational efficiency of the IPEM lies exactly in the fact that the information in the data are condensed into $\hat{\alpha}$ and \hat{P}_{α} . These variables have generally much smaller dimension than N . To maintain the simplicity of the IPEM the initial value $\hat{\theta}^{(0)}$ should be computed from $\hat{\alpha}$ and possibly \hat{P}_{α} . In general terms we can proceed as follows. The estimate $\hat{\alpha}$ is by construction consistent. Hence it will lie close to the set S_2 . Take an arbitrary point $\bar{\alpha}$ in S_2 that is close to $\hat{\alpha}$. We can then regard $\bar{\alpha}$ as a consistent estimate as well. Finally compute $\hat{\theta}^{(0)}$ as

$$\hat{\theta}^{(0)} = \theta(\bar{\alpha})\tag{4.4}$$

cf (2.7). A more specific way to organize such calculations is application dependent.

Example 4.1. (Initial values for GLS). Consider the situation described in Example 1.1. Let the coefficients of A_1 and \bar{B}_1 be determined by

$$\begin{aligned}\bar{B}_1(q^{-1}) &= \bar{B}_2(q^{-1}) + 0(q) \\ \bar{A}_1(q^{-1}) &= \bar{A}_2(q^{-1})\end{aligned}\tag{4.5a}$$

This means that the first $na_1 + nb_1$ values of the impulse responses of the models (1.1a) and (1.1d) will coincide. The relation (4.5a) can be equivalently written as

$$\bar{A}_1(q^{-1}) \hat{B}_2(q^{-1}) - \hat{A}_2(q^{-1}) \bar{B}_1(q^{-1}) = 0(q^{-(na_1 + nb_1 + 1)}).\tag{4.5b}$$

By equating the coefficients of the first $na_1 + nb_1$ powers of q^{-1} in the left-hand side to zero we get the following linear system of equations

$$\begin{aligned} \left(\begin{array}{cc} 0 & 1 \\ \hat{b}_1^{(2)} & -\hat{a}_1^{(2)} \end{array} \right) & \left(\begin{array}{c} 1 \\ -\hat{a}_1^{(1)} \end{array} \right) = \left(\begin{array}{c} 1 \\ -\hat{b}_1^{(2)} \end{array} \right) \\ & \left(\begin{array}{c} \hat{b}_1^{(1)} \\ \hat{b}_{na_1+1}^{(2)} \end{array} \right) \left(\begin{array}{c} -\hat{a}_1^{(2)} \\ \hat{a}_1^{(1)} \end{array} \right) = \left(\begin{array}{c} 1 \\ -\hat{b}_1^{(2)} \end{array} \right) \end{aligned}\tag{4.5c}$$

By convention, the coefficients $\hat{a}_i^{(2)}$ and $\hat{b}_j^{(2)}$ are zero if $i > na_2, j > nb_2$. □

Initial values for the \bar{C} -coefficients may next be determined from one or both of the relations

$$\begin{aligned}\bar{A}_1(q^{-1}) \bar{C}_1(q^{-1}) &\approx \bar{A}_2(q^{-1}) \\ \bar{B}_1(q^{-1}) \bar{C}_1(q^{-1}) &\approx \bar{B}_2(q^{-1}).\end{aligned}\tag{4.5d}$$

Equating the powers of q^{-1} will yield an overdetermined linear system of equations. The least squares solution of this system is a convenient way of computing $\bar{C}_1(q^{-1})$.

Note that the initial values $\bar{A}_1(q^{-1}), \bar{B}_1(q^{-1}), \bar{C}_1(q^{-1})$ computed in this way are consistent by construction. □

To illustrate the properties of the IPEM and the algorithm (4.1) we present some numerical simulations in the following example.

Example 4.2 (Numerical illustrations for GLS). The system

$$y(t) + ay(t-1) = bu(t-1) + \frac{1}{1+cq} \varepsilon(t)$$

$$a = -0.9 \quad b = 1.0 \quad c = -0.9$$

was simulated and its parameters estimated. PC-Matlab (Moler *et al.*, 1987), with the System Identification Toolbox (Ljung, 1986), was used for the computations. Three different methods were used:

- PEM, which in the toolbox is implemented as a Gauss-Newton algorithm for minimizing $\sum_{i=1}^N \varepsilon_i^2(t, \theta)$. The iterations are stopped when the gradient of this criterion has Euclidean norm less than δ . The initial estimate is based on a sophisticated use of the least squares and instrumental variable methods. See Young (1984) or Söderström and Stoica (1989) for descriptions of these methods.
- GLS, which is the traditional method for implementing PEM in the case of generalized least squares model [see Clarke (1967) and Söderström (1974)]. The iterations were stopped when $\|\theta^{(k+1)} - \theta^{(k)}\| \leq \delta(1 + \|\theta^{(k)}\|)$. The initial estimate $\theta^{(0)}$ was formed by making a least squares estimation of a and b . In GLS a and b are estimated alternatively with c . One iteration is here defined as one update of all three parameters.
- IPEM, which was implemented as in (4.1) and initialized as in Example 4.1. To be specific, the initial values of a and b where found from (4.5c) which in this case becomes

$$\begin{pmatrix} 0 & -1 \\ \hat{b}_1^{(2)} & -\hat{a}_1^{(2)} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -\hat{b}_1^{(2)} \\ \hat{b}_1^{(2)} \end{pmatrix}$$

The initial value of c was obtained by equating the coefficients for q^{-1} of the first equation of (4.5d). This gives $a + c = \hat{a}_1^{(2)}$. The iterations were stopped when $\|\theta^{(k+1)} - \theta^{(k)}\| \leq \delta(1 + \|\theta^{(k)}\|)$.

The methods were applied to three different cases

- Case 1: $N = 100$, $u(t)$ white noise of variance 2, $\varepsilon(t)$ white noise of variance 1.
- Case 2: $N = 1000$, $u(t)$ white noise of variance 2, $\varepsilon(t)$ white noise of variance 1.

TABLE 1. CONVERGENCE RESULTS, CASE 1 (w/ WHITE NOISE, $N = 100$)

Method	δ	# iterations	Time (sec)	# kflops	θ		
					(a)	(b)	(c)
PEM	—	1	—	—	-1.0034	0.9689	0.6573
	10^{-2}	5	19	59	-0.8891	0.9917	0.7862
	10^{-3}	7	25	83	-0.8890	0.9912	0.7878
	10^{-4}	9	32	106	-0.8879	0.9911	0.7879
	10^{-5}	10	36	118	-0.8879	0.9911	0.7879
	10^{-6}	12	42	142	-0.8879	0.9911	0.7879
GLS	—	1	—	—	-0.9929	0.9753	0.6605
	10^{-2}	5	12	42	-0.9097	1.0018	0.7623
	10^{-3}	10	24	84	-0.8918	0.9938	0.7840
	10^{-4}	15	35	126	-0.8900	0.9929	0.7862
	10^{-5}	20	47	168	-0.8897	0.9928	0.7865
	10^{-6}	26	61	218	-0.8897	0.9928	0.7865
IPEM	—	1	—	—	-0.8811	0.9895	0.7773
	10^{-2}	2	4	13	-0.8902	0.9943	0.7869
	10^{-3}	3	4	14	-0.8904	0.9944	0.7859
	10^{-4}	4	5	15	-0.8905	0.9945	0.7858
	10^{-5}	6	7	16	-0.8905	0.9945	0.7857
	10^{-6}	8	8	18	-0.8905	0.9945	0.7857

TABLE 2. CONVERGENCE RESULTS, CASE 2 (w/ WHITE NOISE, $N = 1000$)

Method	δ	# iterations	Time (sec)	# kflops	θ		
					(a)	(b)	(c)
PEM	—	1	—	—	-0.7879	1.0317	0.9813
	10^{-2}	4	57	465	-0.8919	0.9912	0.9089
	10^{-3}	5*	99	711	-0.8919	0.9912	0.9089
	10^{-4}	5*	99	711	-0.8919	0.9912	0.9089
	10^{-5}	5*	99	711	-0.8919	0.9912	0.9089
	10^{-6}	5*	99	711	-0.8919	0.9912	0.9089
GLS	—	1	—	—	-0.9852	0.9848	0.7980
	10^{-2}	4	36	332	-0.9263	1.0038	0.8711
	10^{-3}	10	90	831	-0.8976	0.9942	0.9022
	10^{-4}	16	144	1340	-0.8943	0.9926	0.9054
	10^{-5}	22	199	1828	-0.8939	0.9924	0.9057
	10^{-6}	28	256	2327	-0.8939	0.9924	0.9058
IPEM	—	1	—	—	0.8945	0.9926	0.9073
	10^{-2}	2	10	127	0.8947	0.9927	0.9051
	10^{-3}	2	10	127	0.8947	0.9927	0.9051
	10^{-4}	3	11	128	0.8947	0.9926	0.9052
	10^{-5}	4	11	128	0.8947	0.9926	0.9052
	10^{-6}	5	12	129	0.8947	0.9926	0.9052

Case 3: $N = 1000$, $u(t)$ a first order autoregression with pole in $z = 0.8$ and variance 5.56, $\varepsilon(t)$ white noise of variance 1.

The numerical results obtained are given in Tables 1–3. Note that the parameter δ used for the stop criteria has different meaning, for the different methods. For completeness also the initial values are shown, which will give a more complete description of the (practical) convergence properties. In the tables, iteration 1 refers to the initial value.

Since PC-Matlab is used, the computation times shown must be considered with some care (matrix manipulations are fast while user-made loops are comparatively slow). PC-Matlab also provides a counting of the floating point operations. This figure will also give a good indication of the computational load.

The following comments can be made to the results in Table 1. The estimates θ differ a little between the different methods due to different implementations. (Transient effects of filtering etc. differ, for example.) To get convergence to four digits in the result we need to choose $\delta = 10^{-4}$ for PEM, $\delta = 10^{-5}$ for GLS and $\delta = 10^{-5}$ for IPEM. IPEM gives almost the final result after one iteration of (4.1). It is also obvious that IPEM is much more cost-effective in terms of computations than PEM and GLS.

TABLE 3. CONVERGENCE RESULTS, CASE 3 (w/ AUTOREGRESSION, $N = 1000$)

Method	δ	# iterations	Time (sec)	# kflops	θ		
					(a)	(b)	(c)
PEM	—	1	—	—	0.8902	1.0263	0.9028
	10^{-2}	2	33	230	0.8879	1.0225	0.9012
	10^{-3}	3	45	348	0.8876	1.0220	0.9011
	10^{-4}	4	60	466	0.8876	1.0220	0.9011
	10^{-5}	5	73	585	0.8876	1.0220	0.9011
	10^{-6}	7	99	821	0.8876	1.0220	0.9011
GLS	—	1	—	—	0.9413	0.9686	0.7776
	10^{-2}	4	37	332	-0.8923	1.0218	0.9009
	10^{-3}	6	56	499	-0.8991	1.0210	0.9039
	10^{-4}	8	74	665	-0.8988	1.0209	0.9042
	10^{-5}	10	93	831	-0.8987	1.0209	0.9043
	10^{-6}	12	111	997	-0.8987	1.0209	0.9043
IPEM	—	1	—	—	-0.8988	1.0207	-0.9035
	10^{-2}	2	10	127	-0.8988	1.0207	-0.9039
	10^{-3}	2	10	127	-0.8988	1.0207	-0.9039
	10^{-4}	3	11	128	-0.8988	1.0207	-0.9039
	10^{-5}	3	11	128	-0.8988	1.0207	-0.9039
	10^{-6}	4	12	129	-0.8988	1.0207	-0.9039

Compared to the previous results the following comments can be made to Table 2. For four cases (marked *) PEM was stopped after 5 iterations, since the loss function was not reduced further, although the step length in the Newton algorithm was 10 times reduced by 50%. IPEM converges very quickly (The initial values "iteration 1" are very good and essentially one iteration of (4.1) is sufficient.) Note the small computational cost for additional iterations of IPEM (less than 1 klop per iteration compared to 83 klops per iteration for GLS). In this case GLS converges quite slowly.

The results shown in Table 3 are qualitatively similar to those presented in Table 2.

5. Conclusions

An indirect prediction error method has been proposed and analyzed. It is applicable to cases of two nested model structures where a PEM is relatively easy to apply in the larger structure. A simple algorithm was derived for the estimates in the smaller model structure assuming the PEM is applied in the larger structure as a first step.

The proposed method is computationally fast and in that respect superior to a PEM used directly in the smaller structure. The statistical properties are the same if the number of data is large but much less computations are needed with the new method. The indirect PEM approach can be seen as a systematic way of reducing or simplifying models obtained by system identification.

Acknowledgements—The authors are grateful to Professor Michel Gevers for some valuable comments on the manuscript.

References

- Bastin, G., R. R. Bitmead, G. Campion and M. Gevers (1989). Identification of linearly overparametrized non-linear systems. *Proc. 28th IEEE Conf. on Decision and Control*, Tampa, FL, pp. 618–623.
- Clarke, D. W. (1967). Generalized least squares estimation of parameters of a dynamic model. 1st IFAC Symp. on identification in Automatic Control Systems, Prague, paper 3.17.
- Friedlander, B. and B. Porat (1988). Performance analysis of MA parameter estimation algorithms based on high-order moments. *Proc. ICASSP'88*, New York, pp. 2412–2415.
- Dasgupta, S. (1988). Adaptive identification of systems with polynomial parametrizations. *IEEE Trans. Circ. Syst.*, **CAS-35**, 599–603.
- Dasgupta, S., B. D. O. Anderson and R. J. Kave (1988). Identification of physical parameters in structured systems. *Automatica*, **24**, 217–225.
- Jakeman, A. and P. Young (1981). On the decoupling of system and noise model parameter estimation in time-series analysis. *Int. J. Control*, **34**, 423–431.
- Jakeman, A. and P. Young (1983). Advanced methods of recursive time-series analysis. *Int. J. Control*, **37**, 1291–1310.
- Ljung, L. (1976). On the consistency of prediction error identification methods. In R. K. Mehra and D. G. Lainiotis (Eds.), *System Identification—Advances and Case Studies*, Academic Press, New York, pp. 121–164.
- Ljung, L. (1986). *System Identification Toolbox—User's Guide*. Mathworks, Sherborn, MA.
- Ljung, L. (1987). *System Identification—Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ.
- Ljung, L., I. Gustavsson and T. Söderström (1974). Identification of linear multivariable systems operating under linear feedback control. *IEEE Trans. Aut. Control*, **AC-19**, 836–840.
- Moler, C., J. Little and S. Bangert (1987). *PC-Matlab User's Guide*. Mathworks, Sherborn, MA.
- Moses, R. (1986). Optimal approximate stochastic partial realization. In C. J. Byrnes and A. Lindqvist (Eds.), *Modelling, Identification and Robust Control*. North-Holland, Amsterdam, pp. 515–526.
- Porat, B. and B. Friedlander (1986). On the limiting behavior of estimates based on sample covariances. *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Tokyo, pp. 585–588.
- Rosen, Y. and B. Porat (1986). ARMA parameter estimation based on sample covariances, for missing data. *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Tokyo, pp. 209–212.
- Söderström, T. (1974). Convergence properties of the generalized least squares identification method. *Automatica*, **10**, 617–626.
- Söderström, T. (1975). Test of pole-zero cancellation in estimated models. *Automatica*, **11**, 537–541.
- Söderström, T. (1977). On model structure testing in system identification. *Int. J. Control*, **26**, 1–18.
- Söderström, T. and P. Stoica (1989). *System Identification*, Prentice-Hall, Hemel Hempstead, U.K.
- Stoica, P. (1976). The repeated least squares identification method. *Journal A*, **17**, 151–156.
- Stoica, P. and T. Söderström (1982). On the parsimony principle. *Int. J. Control*, **36**, 409–418.
- Stoica, P. and T. Söderström (1989). On reparametrization of loss functions used in estimation, and the invariance principle. *Signal Processing*, **17**, pp. 383–387.
- Stoica, P., B. Friedlander and T. Söderström (1985a). The parsimony principle for a class of model structures. *IEEE Trans. Aut. Control*, **AC-30**, 597–600.
- Stoica, P., B. Friedlander and T. Söderström (1985b). Large sample estimation of the AR parameters of an ARMA model. *IEEE Trans. Aut. Control*, **AC-30**, 891–893.
- Stoica, P., B. Friedlander and T. Söderström (1987). Approximate maximum-likelihood approach to ARMA spectral estimation. *Int. J. Control*, **45**, 1281–1310.
- Wahlberg, B. (1986). On model reduction in system identification. *Proc. Amer. Control Conf.*, Seattle, WA, pp. 1260–1266.
- Wahlberg, B. (1987). On the identification and approximation of linear systems. Doctoral Dissertation (no. 163), Department of Electrical Engineering, Linköping University, Linköping, Sweden.
- Wahlberg, B. (1989). Estimation of ARMA models via high order autoregressive approximations. *J. Time Series Anal.*, **10**, 283–299.
- Young, P. (1984). *Recursive Estimation and Time-Series Analysis*. Springer, Berlin.

Brief Paper

Comments on 2-D Descriptor Systems*

STEPHEN L. CAMPBELL†

Key Words—Boundary-value problem; difference equations; discrete systems; multidimensional systems; linear systems.

Abstract—This paper discusses general 2-D descriptor systems of the form $Ex_{i+1,j+1} = Ax_{i,j} + Bx_{i+1,j} + Cx_{i,j+1} + D_0u_{i,j} + D_1u_{i+1,j} + D_2u_{i,j+1}$. Solution formula and structural forms are developed for several large classes of 2-D descriptor systems. The idea of a recursive chain is introduced. It is shown that these systems include several types of behavior not discussed elsewhere in the literature.

1 Introduction

RECENTLY THERE has been increasing interest in 2-D (and the more general m -D) systems because of their many applications in the numerical analysis of PDEs, image processing, signal processing, and as models for other discrete processes. Singular (or descriptor) systems, more frequently called differential-algebraic (DAE) in the mechanics and numerical literature (Brenan *et al.*, 1989), have become increasingly important for 1-D systems. Singular 2-D systems should also become increasingly important. This paper will make several observations about singular 2-D systems. We are particularly interested in determining properties of 2-D systems which are *not* seen with singular 1-D systems and have not been previously discussed in the literature.

The system of interest is

$$Ex_{i+1,j+1} = Ax_{i,j} + Bx_{i+1,j} + Cx_{i,j+1} + \sum_{\alpha=0,1,2} D_\alpha u_{i+\alpha,j+\alpha} \quad (1)$$

$$v_{i,j} = Gx_{i,j} + Hu_{i,j}, \quad i \geq 0, j \geq 0 \quad (2)$$

where $x_{i,j}$ has values in \mathcal{R}^n (or \mathcal{C}^n), E is a singular square matrix, the rest of the matrices are conformable in size, and $I_{i,j}$ is a finite index set which depends on (i,j) . For this paper it suffices to consider

$$Ex_{i+1,j+1} = Ax_{i,j} + Bx_{i+1,j} + Cx_{i,j+1} + f_{i,j}, \quad i \geq 0, j \geq 0 \quad (3)$$

Let \mathcal{J}^* be integer 2-tuples (i,j) such that $i \geq 0, j \geq 0$. The boundary of \mathcal{J}^* , denoted $\partial\mathcal{J}^*$, is those $(i,j) \in \mathcal{J}^*$ with at least one of i, j equal to 0. For any set Σ , $\mathcal{F}_n(\Sigma)$ is the vector space of all \mathcal{R}^n (or \mathcal{C}^n) valued functions defined on Σ . Thus we can talk of $f \in \mathcal{F}_n(\mathcal{J}^*)$, or of a sequence $\{f_{i,j}, i \geq 0, j \geq 0\}$. Given $x \in \mathcal{F}_n(\mathcal{J}^*)$, its *boundary values* are its values for $(i,j) \in \partial\mathcal{J}^*$. The system (3) is called *solvable* if for any $f \in \mathcal{F}_n(\mathcal{J}^*)$, there is a solution to (3) in $\mathcal{F}_n(\mathcal{J}^*)$ which is

uniquely determined by its boundary values. Given $f \in \mathcal{F}_n(\mathcal{J}^*)$, a boundary value $\{x_{i,j}, i \geq 0 \text{ or } j \geq 0\}$ is *consistent* for (3) if there is a solution to (3) with this boundary value.

The system (3) with E nonsingular has been studied in many papers. As in Fornasini and Marchesini (1978), we prefer (3) to other models, such as the Roesser. Firstly, (3) permits smaller sized arrays. Secondly, (3) allows the use of general similarity transformations in both analysis and computation. Finally, as discussed in the section on alternative forms, other model forms disguise some of the difficulties present. The singular case, under different assumptions from ours, is also studied in Kaczorek (1988) and Lewis and Mertzios (1988).

2 Solutions

System (3) is *regular* if $\det(zI - A - zB - uC)$ is nonsingular for some pair z, u , and *simply regular* if either $zE - B$ or $zE - C$ is a regular pencil. Simply regular implies regular. If A is nonsingular, but the other coefficients are zero, then the system is regular but not simply regular. A more interesting example is when $E = 0$ and B, C are both singular but $B + C$ is nonsingular. Such systems, which include the usual Roesser's model (Kaczorek, 1985), are not simply regular. Note that solvability implies regularity.

In order to examine regular systems, we need to review a few concepts from Campbell (1980). Recall that every square matrix A has a Drazin inverse denoted A^D . There are many ways to express the solution of a 1-D descriptor difference equation, including the use of shuffle-like algorithms, but we shall use the Drazin inverse notation because it is convenient. The numerical implementation of this notation is discussed in Wilkinson (1982). The *index* of a matrix is the size of the largest nilpotent block in its Jordan canonical form. Then from Campbell (1989) we have

Theorem 1. Suppose that E, A is a regular pencil. Then

$$Ex_{k+1} = Ax_k + f_k, \quad k \geq 0 \quad (4)$$

is solvable and the general solution is given by

$$x_k = A^D P q + E^D \sum_{i=0}^{k-1} A^i + (I - E^D A^D) \sum_{i=0}^{k-1} E A^i f_{k-i} \quad (5)$$

where $f_i = (zI - A)^{-1} (I - E^D A^D) f_i$, $A^D = (zI - A)^{-1} A$, $A^D A = I - E^D A$, $E^D E = I - EA^D$, $P = EE^D$, q is an arbitrary vector, ν is the index of E and λ is a scalar such that $\lambda I - A$ is nonsingular. The projection P and $E - A^D A$ are independent of λ .

Assume then that E, C is a regular pencil. A similar discussion applies if E, B is regular. Treat j as fixed. If the sequence $x_{i,j}$ is considered known, then (3) is a difference equation

$$E x_{i+1,j+1} = C x_{i,j+1} + [A x_{i,j} + B x_{i+1,j} + f_{i,j}], \quad i \geq 0 \quad (6)$$

for $x_{i,j+1}$, with the terms in square brackets known. Since E, C is a regular pencil, we may apply Theorem 1 and determine $x_{i,j+1}$. Several important consequences of this approach are summarized in the next proposition.

* Received 24 October 1989; revised 20 March 1990; received in final form 14 April 1990. The original version of this paper was presented at the IFAC Workshop on System Structure and Control which was held in Prague, Czechoslovakia during September, 1989. The published Proceedings of this IFAC Meeting may be ordered from Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor V. Kučera under the direction of Editor H. Kwakernaak.

† Department of Mathematics & Center for Research in Scientific Computation, Box 8205, North Carolina State University Raleigh, NC 27695-8205 USA.

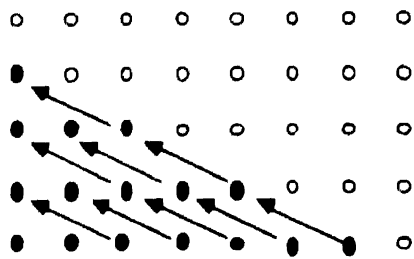


FIG. 1. Simply regular with $\hat{v} = 2$.

Proposition 1 Suppose that E, C is a regular pencil. Then (3) is solvable. Let $\hat{v} = \text{Index}(\hat{E})$, $\hat{P} = \hat{E}^P \hat{E}$, where $\hat{E} = (\lambda E - C)^{-1} E$. Set $\hat{v} = \hat{v} + 1$. Then

- 1. The boundary values $x_{i,0}$ may be taken arbitrary.
- 2. The boundary values $Px_{0,j}$, $j \geq 1$, are arbitrary.
- 3. The boundary values $(I - P)x_{0,j}$, $j \geq 1$, are determined by $\{x_{i,0} | 0 \leq i \leq j\hat{v}\}$ and $\{f_{i,j} | 0 \leq i \leq j-1, 0 \leq i \leq (j-1)\hat{v}\}$.
- 4. For $i \geq 1$ the value of $x_{i,j}$ is determined from the values of $x_{i,j}$ for $0 \leq i \leq i-1$ and $x_{i,j-1}$, $f_{i,j-1}$ for $0 \leq i \leq (j-1)\hat{v}$.

This computation can be carried out in an increasingly concurrent fashion. The most interesting case is when $\hat{v} \geq 1$. Given the boundary conditions, the computation proceeds as follows.

- 1. We begin computing $x_{1,1}, x_{2,1}, x_{3,1}, \dots$.
- 2. As soon as x_{v+1} has been computed we can begin to concurrently compute $x_{1,2}, x_{2,2}, x_{3,2}, \dots$.
- 3. Once $x_{1,v}, \dots, x_{v,v}$ are computed, we begin computing $x_{i,j+1}$ for $i = 1, 2, \dots$.

This pattern is illustrated in Fig. 1 for the $\hat{v} = 2$ case. These results appear to contradict the characterization of solvability in Kaczorek (1988). However, there is an implicit assumption in Kaczorek (1988) that the value of $x_{i,j}$ can be determined without knowing any values of either $x_{r,i}$ or $f_{r,j}$ with $r < i$ or $s < j$. In the case of a simply regular system this forces \hat{v} to be zero or one. While useful, there is less of a logical reason to make this assumption than in the 1-D case since i, j may be spatial rather than temporal variables.

Example 1 Let

$$E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad A = C = 0.$$

This problem is simply regular with $\hat{v} = 2$ but it does not satisfy the rank conditions in Kaczorek (1988) since, in the notation of that work, we have $A_{22} = 0$ but $B_2 \neq 0$.

A convenient way to visualize the computation of a solution of (3) is by picturing the region where at least part of $f_{i,j}$ (or $x_{i,j}$) must be known in order to compute $x_{i,j}$. For example, Fig. 1 could be pictured as in Fig. 2.

We now construct examples of solvable 2-D systems with more interesting computational configurations. Consider the 2-D system

$$E_1 x_{i+1,j+1} = B_1 x_{i,j} + C_1 y_{i,j+1} + f_{i,j} \tag{7}$$

$$E_2 y_{i+1,j+1} = C_2 y_{i,j+1} + g_{i,j} \tag{8}$$

Let

so that (7), (8) is a 2-D descriptor system in $z_{i,j}$ in the form of (3).



FIG. 2. Simply regular with $\hat{v} = 2$.



FIG. 3. Example 2.

Example 2. Suppose in (7), (8), that $C_1 = 0$ and

$$B_1 = C_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad E_1 = E_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Then the data required to compute $z_{i,j}$ is pictured in Fig. 3.

Example 3. Suppose we change C_1 in Example 2 to $C_1 = C_2 = I$. Then the region is given in Fig. 4. That is, we need values of $y_{i,j}$ in the shaded region of Fig. 4 to determine $x_{i,j}$ and these values in turn require information from the cross-hatched area.

If we view (6) as having input $f_{i,j}$ and output $x_{i,j}$, then the output mask (region where data is needed) for Fig. 4 does not appear to be what Bose (1982) calls recursively computable. Yet the difference equations are solvable. The reason for this is that different masks are used to compute different parts of $x_{i,j}$ and one mask may have to wait until another mask has finished computing needed values. This is discussed more carefully in the next section.

3. Recursive chains

The previous examples are special cases of a wide class of solvable 2-D systems which are not recursive in the usual sense. To discuss these systems we need to extend some previous terminology.

For $n \times n$ matrices E, B, C, A we define the operator \mathcal{L} of $\mathcal{F}_n(\mathcal{Z}^+)$ into itself by

$$(\mathcal{L}[x])_{i,j} = Ex_{i+1,j+1} + Bx_{i,j} + Cx_{i,j+1} + Ax_{i,j} \tag{9}$$

Note that $\mathcal{L}[x] = f$ is solvable if the range of \mathcal{L} is $\mathcal{F}_n(\mathcal{Z}^+)$ and $\mathcal{L}[z] = 0$, $z|_{\mathcal{Z}^+} = 0$, implies $z = 0$.

\mathcal{L} is a *recursive operator* if \mathcal{L} is solvable and the solution of $\mathcal{L}[x] = f$ can be computed recursively (Bose, 1982) given f and consistent boundary conditions. That is, there is a suitable mask I_n so that:

$$x_{i,j} = \sum_{(r,s) \in I_n} [\alpha_{r,s} x_{i-r,j-s} + \beta_{r,s} f_{i-r,j-s}] \tag{10}$$

Example 3 is not recursive in the sense of Bose (1982).

Proposition 2. The following cases are all recursive, and hence solvable, operators.

- I. E, B or E, C a regular pencil. This includes E nonsingular.
- II. $B = C = 0$ and E, A a regular pencil.
- III. $E = C = 0$ and C, A a regular pencil.
- IV. $E = B = 0$ and B, A a regular pencil.

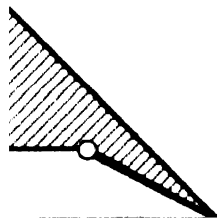


FIG. 4. Example 3.

We denote a sequence from $\mathcal{B}_n(\mathcal{X}^*)$ by $z^{[i]}$. A function $F(z^{[1]}, \dots, z^{[p]})$ from $\prod_{i=1}^p \mathcal{B}_n(\mathcal{X}^*)$ into \mathcal{B}^* has finite domain if there exists an integer ρ such that the value of F depends only on the numbers $(z_k^{[i]}, 1 \leq k \leq r, 0 \leq i \leq \rho, 0 \leq j \leq \rho)$. We define the system (3) to be an r -stage chain if it is in the form

$$\begin{aligned} \mathcal{L}_1[x^{[1]}] &= \tilde{F}_1(x^{[2]}, \dots, x^{[r]}, f) \\ \mathcal{L}_2[x^{[2]}] &= \tilde{F}_2(x^{[3]}, \dots, x^{[r]}, f) \\ &\vdots \\ \mathcal{L}_{r-1}[x^{[r-1]}] &= \tilde{F}_{r-1}(x^{[r]}, f) \\ \mathcal{L}_r[x^{[r]}] &= \tilde{F}_r(f) \end{aligned} \quad (11)$$

and each \tilde{F}_i is a sequence of vector valued functions with finite domains. If all the operators \mathcal{L}_i in (11) are solvable, then the system (11) is solvable. If each \mathcal{L}_i is recursive, then (11) will be called r -stage recursive, or a recursive chain. Example 3 was two-stage recursive but not recursive.

An r -stage recursive chain can be thought of as having r masks each operating on a different portion of the state vector with these masks arranged in a recursive manner.

For the nonsingular case, a sufficient condition (Fornasini and Marchesini, 1979, 1980; Pandolfi, 1984) for bounded input bounded output (BIBO) of (3) is that

$$p(z_1, z_2) = \det(E - z_1 B - z_2 C - z_1 z_2 A) \quad (12)$$

have no zeros for which $|z_1| \leq 1$ and $|z_2| \leq 1$.

Example 4. This is an example of case II of Proposition 2. Let

$$E = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = C = 0, \quad A = I.$$

Then $p(z_1, z_2) = z_1^2 z_2^2$ so there are zeros of p for which $|z_1| \leq 1, |z_2| \leq 1$. However the system will be BIBO as long as the domain and range have the same norm since the solution is $x_{i,j} = f_{i,j} - E f_{i,j+1,j+1}$.

4. Alternative forms

Many alternative forms for 2-D systems have been discussed in the literature. There are certain difficulties in altering forms with singular 2-D systems. To illustrate, we take an example from Fornasini and Marchesini (1978) that is typical in that it utilizes a new state made up of shifted values of the previous state. Consider

$$E x_{i,j+1,j+1} = A_1 \tilde{x}_{i,j+1} + A_2 \tilde{x}_{i,j+1} + \tilde{A}_0 \tilde{x}_{i,j} + \tilde{B} u_{i,j} \quad (13)$$

$$y_{i,j} = \tilde{C} \tilde{x}_{i,j} \quad (14)$$

This same input-output relationship can be written as

$$E w_{i,j+1,j+1} = A_1 w_{i,j+1} + A_2 w_{i,j+1} + B_1 u_{i,j+1} + B_2 u_{i,j+1} \quad (15)$$

$$y_{i,j} = C w_{i,j} \quad (16)$$

by taking

$$\begin{aligned} E &= \begin{bmatrix} \tilde{E} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, \quad w_{i,j} = \begin{bmatrix} x_{i,j} \\ r_{i,j} \\ y_{i,j} \end{bmatrix}, \\ A_1 &= \begin{bmatrix} \tilde{A}_1 & 0 & 0 \\ I & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} \tilde{A}_2 & \tilde{A}_0 & B \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\ B_1 &= \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad C = [\tilde{C} \quad 0 \quad 0] \end{aligned}$$

where $r_{i,j+1,j+1} = x_{i+1,j+1}$, $\tilde{x}_{i+1,j+1} = u_{i,j+1}$. The new system (15), (16) is simpler, since \tilde{A}_0 in (13) has been eliminated. However, what has happened to the boundary values? Going from (13), (14) to (15), (16) is the

same as taking

$$w_{i,j} = \begin{bmatrix} x_{i,j} \\ r_{i,j+1} \\ y_{i,j} \end{bmatrix} \quad (17)$$

Assume we have the boundary values $x_{i,0}$, $x_{0,j}$ and forcing function $u_{i,j}$. But $w_{i,j}$ in (17) is not defined for $j = 0$. Thus we must consider (15) as defined for $i \geq 0, j \geq 1$, and the needed boundary values are $w_{i,1}$ and $(w_{0,j}, j \geq 1)$. We have the needed values of $r_{i,j}$, $x_{i,j}$. However, the values of $x_{i,1}$ will have to be computed. If E is singular, the boundary conditions are restricted to an affine subspace. But now the computation of the new boundary conditions is almost equivalent to solving the original difference equation. This problem of determining the consistent boundary values for alternative forms based on shifted state values is one of the reasons we prefer to work with the original 2-D system (3).

5. Infinite dimensionality

The 2-D system (3) is the discrete analogue of a partial differential equation (PDE). Many of the differences between 2-D and 1-D systems are due to the fact that, like PDEs, the system is intrinsically infinite dimensional.

5.1. Waves. As with PDEs it is natural to look for traveling wave solutions of (3). They are easier to compute and may have physical interpretations.

Let m, n be nonnegative but fixed integers. We could allow them to be general scalars but it will simplify the exposition if we assume they are integers. We seek solutions of the associated homogeneous equation for (3) in the form

$$x_{i,j} = \phi(mi + nj) \quad (18)$$

The vector function ϕ need only be defined on the numbers given by $\{mi + nj, i \geq 0, j \geq 0\}$ where m, n are any two real numbers. However, we shall assume ϕ is defined for all integers. If (18) is a solution of (3) with $f = 0$, then

$$I \phi(z + m + n) = A \phi(z) + B \phi(z + m) + C \phi(z + n) \quad (19)$$

where $z = mi + nj$. But (19) is a difference equation for ϕ . Introducing variables $\psi_1(z) = \phi(z + 1)$, $\psi_2(z + 1) = \psi_1(z)$ with $r = m + n$, but $m \neq n$, gives the 1-D description system

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ & & 1 \end{bmatrix} \begin{bmatrix} \phi(v + 1) \\ \psi_1(v + 1) \\ \psi_2(v + 1) \end{bmatrix} = \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & 0 \\ A & * & * \end{bmatrix} \begin{bmatrix} \phi(v) \\ \psi_1(v) \\ \psi_2(v) \end{bmatrix} \quad (20)$$

Equation (20) will have nontrivial solutions for ϕ provided that $\det(\lambda^m I - \lambda^n B - \lambda^n C - A)$ is nonconstant.

Example 5. $\phi(z) = a^z$ with $m = n = 1$ gives the "traveling wave" solution a^{i+j} of $\alpha x_{i,j+1,j+1} + \beta x_{i,j+1} + \gamma x_{i,j+1} + \delta x_{i,j} = 0$ if $a = (\alpha + \gamma) / [\beta + \gamma + (\beta + \gamma + 4\alpha\delta)^{1/2}]$.

5.2. Operator formulation. As with PDEs it can be helpful to view the difference equation (3) as an operator difference equation. When we do this we see that some of our previous examples exhibit behavior which is intrinsically infinite dimensional in character.

We shall take one of i, j as the independent variable. Let \mathcal{C}^p be the usual space of p -summable sequences of real or complex numbers with norm $\|\cdot\|_p$. If $p = \infty$, then this is the sup norm. Let \mathcal{C}_n^p be the vector space of all sequences of n -dimensional vectors. Then \mathcal{C}_n^p is those $x \in \mathcal{C}_n^p$ such that $\{\|x_k\|_p\} \in \mathcal{C}^p$ for some fixed norm $\|\cdot\|$. We then define $\|x\|_p = \{\|x_k\|_p\}_p$. Now let $z \in \mathcal{C}_n^p$ be the column vector created by listing $x_{0,0}, x_{1,0}, \dots$ in order. Then (3) may be written as

$$E z_{j+1} = A z_j + f_j \quad (21)$$

where

$$E = \begin{bmatrix} -C & E & 0 \\ 0 & -C & I \\ 0 & 0 & -C \end{bmatrix}, \quad A = \begin{bmatrix} A & B & 0 \\ 0 & A & B \\ 0 & 0 & A \end{bmatrix}$$

and $\hat{f}_j = \{f_{0,j}, f_{1,j}, \dots\}$. Since we are now dealing with infinite matrices it is possible for \hat{E} to be onto but not invertible. In fact,

Proposition 3 The matrix \hat{E} in (21) is onto as a linear transformation of ℓ_∞ into itself if and only if the pencil E, C is regular.

If E is onto, then z_0 can be taken arbitrary in (21) and the part of z_{j+1} in the nullspace of E is uniquely determined by its $(0, j+1)$ component, which are the additional consistent boundary conditions in addition to z_0 . Thus these systems are solvable. Notice that the pencil \hat{E}, A need not be regular in that $\lambda\hat{E} + A$ need not have an inverse for any λ . Example 1 has \hat{E} onto but \hat{E} has a nullspace and $A = 0$.

Some care must be taken when working with infinite matrices which are not operators since multiplication is not associative and matrices with nullspaces can also have inverses.

Example 3, on the other hand is not simply regular, so that E is not onto. For Example 4, \hat{E} is nilpotent and $A = I$ so that we can use the algebra used to prove Theorem 1 to show that $z_j \approx \hat{f}_j + E\hat{f}_{j+1}$. Here the row finiteness (finite number of nonzero entries in each row) allows us to associate the needed matrix products.

6. Conclusions

The usual definition of recursiveness has been shown to be inadequate to describe solvable 2-D descriptor systems. The concept of a recursive chain has been introduced and examples given. Motivated by the theory for PDEs, several new types of behavior have been pointed out for 2-D systems.

Acknowledgement—This research was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-87-0051.

References

- Bose, N. K. (1982). *Applied Multidimensional Systems Theory*. Van Nostrand, New York.
- Brenan, K. E., S. L. Campbell and L. R. Petzold (1989). *The Numerical Solution of Initial Value Problems in Differential Algebraic Equations*. Elsevier, New York.
- Campbell, S. L. (1980). *Singular Systems of Differential Equations*. Pitman, London.
- Fornasini, E. and G. Marchesini (1978). Doubly indexed dynamical systems: state-space models and structural properties. *Math. Syst. Theory*, **12**, 59–72.
- Fornasini, E. and G. Marchesini (1979). On the internal stability of two-dimensional filter. *IEEE Trans. Aut. Control*, **AC-24**, 129–130.
- Fornasini, E. and G. Marchesini (1980). Stability analysis of 2-D systems. *IEEE Trans. Circ. Syst.*, **CAS-27**, 1210–1217.
- Kaczorek, T. (1985). *Two-Dimensional Linear Systems*. Springer, Berlin.
- Kaczorek, T. (1988). Singular general model of 2-D systems and its solution. *IEEE Trans. Aut. Control*, **AC-33**, 1060–1061.
- Lewis, F. L. and B. G. Mertzios (1988). On the analysis of two-dimensional discrete singular systems. *Circ. Syst. Signal Process.* (to appear).
- Pandolfi, L. (1984). Exponential stability of 2-D systems. *Syst. Control Lett.* **4**, 381–385.
- Wilkinson, J. (1982). The practical significance of the Drazin inverse. In S. L. Campbell (Ed.), *Recent Applications of Generalized Inverses*, pp. 82–99. Pitman, London.

Brief Paper

Solutions to the H_∞ General Distance Problem which Minimize an Entropy Integral*

D. MUSTAFA,† K. GLOVER‡ and D. J. N. LIMBEER§

Key Words—Multivariable control systems; control systems synthesis; robust control; suboptimal control; optimal control; state space

Abstract—We pose, and solve, the problem of minimizing the entropy of an H_∞ -norm bounded and stabilized closed-loop. Solution proceeds via the equivalent error system distance problem. The central member of the admissible class is shown to minimize the entropy at infinity, and in that case an explicit state-space formula is derived for the minimum value of the entropy. Links between entropy H_2 -norms and H_2 -optimal control are given.

Notation

The notation

$$X = Ric \begin{bmatrix} A & B \\ C & A^T \end{bmatrix} = BB^T - C^T C^{-1} A^T$$

means that $X = X^T$ is the stabilizing solution of the algebraic Riccati equation

$$XA + A^T X - XBB^T X + C^T C = 0$$

State-space realizations will be written

$$\begin{bmatrix} \dot{x} \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \quad \dot{x} = D + C(M - A)^{-1}B$$

A square transfer function matrix $G(s)$ is said to be *all-pass* if

$$G^*(j\omega)G(j\omega) = I \quad \forall \omega$$

The Laplace transform variable s will be suppressed for notational simplicity. All transfer function matrices are taken to have real-rational elements, all constant matrices are taken to have real elements. Other notational conventions are listed below.

- \mathcal{U} The open right half plane.
- \mathcal{R} (Prefix) real-rational.
- \mathcal{RH}_∞ Hardy space of real-rational transfer function matrices.
- \mathcal{RH}_2 Hardy space of real-rational transfer function matrices.
- $\lambda_i[G]$ The i th eigenvalue of G .

* Received 11 April 1989; revised 13 November 1989; received in final form 26 February 1990. A preliminary version of this paper, entitled "Controllers which satisfy a closed-loop H_∞ -norm bound and maximize an entropy integral", was presented by the first two authors at the 1988 Conference on Decision and Control, Austin, TX, U.S.A. This paper was recommended for publication in revised form by Associate Editor H. Kimura under the direction of Editor H. Kwakernaak.

† Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A. Author to whom all correspondence should be addressed.

‡ Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.

§ Department of Electrical Engineering, Imperial College, Exhibition Road, London SW7 2BT, U.K.

- G^* $= G^T(s)^T$, the parahermitian conjugate of G .
- $\sigma_i(G)$ $= \lambda_i^{1/2}[G^*G]$, the i th singular value of G .
- $\|G\|_2$ $= \{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \text{trace}[G^*(j\omega)G(j\omega)] d\omega \}^{1/2}$, the H_2 -norm.
- $\|G\|_\infty$ $= \sup_{\omega} \sigma_i(G(j\omega))$, the H_∞ -norm.
- $\mathcal{B}(H)$ Open unit ball in \mathcal{H} , i.e. all $\Phi \in \mathcal{H}$ such that $\|\Phi\| < 1$.
- $M > 0$ $M = M^T$ is positive definite.
- $M \geq 0$ $M = M^T$ is positive semi-definite.
- $\mathcal{U}(s)$ Denotes minimum entropy.
- \mathcal{U}_∞ Denotes H_∞ optimal.
- \mathcal{U}_2 Denotes H_2 optimal.
- $\star(P, K)$ Lower linear fractional map of P and K .

1. Introduction

THIS PAPER is concerned with *sub-optimal* H_∞ control with a *minimum entropy* criterion. The theory of *optimal* H_∞ control has received much attention over recent years, for full details the interested reader is referred to Francis (1987) and the references therein. Figure 1 illustrates the usual configuration, where the "standard plant" P , consisting of the actual plant, suitable weighting functions and interconnections, maps exogenous inputs w and control inputs u to controlled outputs z and measured outputs y . As is usual in H_∞ control problems, we assume that

$$P \in \mathcal{RH}_\infty$$

The closed-loop transfer function matrix G from w to z is given by the linear fractional map

$$G = \star(P, K) = (P_{22} + P_{21}K(I - P_{11}K)^{-1}P_{12})$$

of the appropriately partitioned standard plant P and the controller K . The optimal H_∞ control problem is to find a stabilizing controller K_{opt} which minimizes the H_∞ norm of this transfer function, i.e. K_{opt} satisfies

$$\inf_{\{K \in \mathcal{RH}_\infty \mid K \text{ stabilizes } P\}} \|\star(P, K)\|_\infty = \gamma \quad (1)$$

Motivated by the belief that optimal H_∞ control is not always appropriate we consider here the sub-optimal problem obtained by relaxing the infimum in (1) to the upper bound

$$\|\star(P, K)\|_\infty \leq \gamma$$

where $\gamma > \gamma_{\text{opt}}$. In general, there is a class of controllers which satisfy this bound, such nonuniqueness is dealt with in this paper by specifying that the *entropy* of the closed-loop transfer function must be *minimized*.

The entropy is defined as follows. For any transfer function matrix G which satisfies $\|G\|_\infty \leq \gamma$, the entropy of

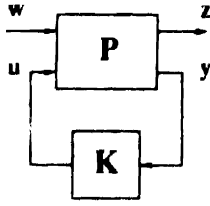


FIG. 1. The standard H_2 control configuration

G , at a point $s_0 \in \mathbb{C}_+$, is defined by

$$I(G; \gamma, s_0) = -\frac{\gamma^2}{2\pi} \int_{-\infty}^{\infty} \ln |\det(I - \gamma^{-2} G^*(j\omega)G(j\omega))| \times \left[\frac{\operatorname{Re} s_0}{|s_0 - j\omega|} \right]^2 d\omega. \quad (2)$$

This definition is equivalent to that of Arov and Krein (1981, 1983), except for an extra $\operatorname{Re} s_0$ term which ensures a nonzero entropy in the case $s_0 \rightarrow \infty$. It is easily seen that the entropy is well-defined (since $\|G\|_\infty < \gamma$ implies $0 < I - \gamma^{-2} G^*(j\omega)G(j\omega) \approx I$) and non-negative; that $I(G, \gamma; s_0) = 0$ if and only if $G = 0$; and that the entropy of G is invariant under unitary scaling of G .

Our minimum entropy H_2 control problem is then:

Find, out of all controllers K which stabilize P and satisfy

$$\|\mathcal{F}(P, K)\|_\infty < \gamma, \quad (3)$$

a K which minimizes $I(\mathcal{F}(P, K); \gamma; s_0)$, the closed-loop entropy at a point $s_0 \in \mathbb{C}_+$.

Minimum entropy has been studied in a wide variety of contexts; its applications to extension problems (Arov and Krein, 1981, 1983) and to contractive interpolants (Gohberg *et al.*, 1988; Dym and Gohberg, 1986, 1988 and see Remark 4) are pertinent here—we use an adaptation of the method of Arov and Krein (1983) in order to obtain a self-contained derivation. The use of minimum entropy in H_2 control has been considered in Limebeer and Hung (1987) for the ‘one-block case’, where both P_{12} and P_{21} are assumed square.

As with optimal H_2 problems (see Francis, 1987) we approach the problem by reducing our original problem to a ‘distance problem’. To do this, use the parametrization of all stabilizing controllers of Youla *et al.* (1976) and Kucera (1979) to reduce (3) to the equivalent model-matching problem of finding $\hat{Q} \in \mathcal{RH}_2$ such that

$$\|T_1 + T_2 \hat{Q} T_3\|_\infty < \gamma, \quad (4)$$

and then exploit the unitary invariance of the H_2 -norm to reduce (4) to the distance problem

$$\left\| \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + \hat{Q} \end{bmatrix} \right\|_\infty < \gamma, \quad \hat{Q} \in \mathcal{RH}_2,$$

where

$$R = \begin{matrix} & \xleftrightarrow{m_1 \quad p_2} & \xleftrightarrow{m_2} \\ \begin{matrix} p_1 - m_2 \downarrow \\ p_2 \downarrow \end{matrix} & \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \end{matrix}$$

is anticausal and is known in terms of the standard plant P .

If we define the error system E by

$$E := \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + \hat{Q} \end{bmatrix} \quad (5)$$

we know that $I(\mathcal{F}(P, K); \gamma; s_0) = I(E; \gamma; s_0)$ because entropy is unitarily invariant (see above). Hence, the closed-loop transfer function $\mathcal{F}(P, K)$ and the error system E have the same entropy. This allows us to solve our original closed-loop problem (3) by solving the following error system distance problem:

Find the error system E as defined in equation (5), which minimizes $I(E; \gamma; s_0)$ over those E which satisfy $\|E\|_\infty < \gamma$.

Two approaches to the solution of the minimum entropy H_2 control problem are evident: direct solution for the controller or solution via the equivalent error system distance problem. Direct solutions to H_2 control problems have only recently appeared (see Glover and Doyle, 1988; Doyle *et al.*, 1989) and it is indeed possible to directly derive the minimum entropy H_2 controller. Such a direct solution is carried out in Glover and Mustafa (1989). However, we present here an alternative derivation via the distance problem, which, given the longstanding role of the distance problem in H_2 control, is of independent interest.

The arrangement of the paper is as follows. In the next section we briefly motivate minimum entropy H_2 control. Section 3 contains the main results. By parametrizing all solutions of the error system distance problem we are able to derive the unique minimum entropy solution together with a value for its entropy. This is firstly done for the general case of entropy at any $s_0 \in \mathbb{C}_+$ and then for the important special case of $s_0 \rightarrow \infty$. This latter case yields particularly explicit and appealing results, and state-space formulae are given. Section 4 contains the concluding remarks.

2. Motivation

Here we briefly state some relevant background details. Recall that we want our controller to stabilize P and keep the H_2 -norm of the closed-loop $\mathcal{F}(P, K)$ below a level γ (where $\gamma > \gamma_{\text{opt}}$). Such control problems lead to a class of admissible controllers. Our approach is to select the admissible controller which minimizes the closed-loop entropy. The closed-loop entropy (2) is a useful measure of how close $G = \mathcal{F}(P, K)$ is to the upper bound γ on $\sigma_1(G(j\omega))$. In particular, if we rewrite the entropy as

$$I(G; \gamma, s_0) = -\frac{\gamma^2}{2\pi} \int_{-\infty}^{\infty} \sum_i \ln |1 - \gamma^{-2} \sigma_i^2(G(j\omega))| \times \left[\frac{\operatorname{Re} s_0}{|s_0 - j\omega|} \right]^2 d\omega, \quad (6)$$

and $\sigma_1^2(G(j\omega)) \rightarrow \gamma^2 - \epsilon^2$ for some frequency range $\omega_1 < \omega < \omega_2$, then $I(G; \gamma, s_0) \rightarrow \infty$ as $\epsilon \rightarrow 0$. Also, it is clear from (6) that all the singular values $\sigma_i(G)$ of G are included, unlike the H_2 -norm which depends only on the largest singular value $\sigma_1(G)$.

The term $[(\operatorname{Re} s_0)/|s_0 - j\omega|]^2$ in the entropy integral is a frequency weighting, with a shape dependent on the position of the point s_0 in the right half plane. In order to obtain real-rational controllers, s_0 should be a real number; allowing $s_0 \rightarrow \infty$ makes the frequency weighting equal to unity for all frequencies, a notable special case we will return to later.

An interesting link with the H_2 -norm is provided by the following lemma.

Lemma 1. Let G be a transfer function matrix which satisfies $\|G\|_\infty < \gamma$. Then

- (a) (i) $I(G; \gamma; s_0) \leq \|G(s)(\operatorname{Re} s_0)/(s_0 + s)\|_2^2$,
(ii) $I(G; \gamma; s_0) = \|G(s)(\operatorname{Re} s_0)/(s_0 + s)\|_2^2 + O(\gamma^{-2})$.
- (b) If G is strictly proper then
(i) $I(G; \gamma; \infty) \leq \|G\|_2^2$,
(ii) $I(G; \gamma; \infty) = \|G\|_2^2 + O(\gamma^{-2})$.

Equality in both (a)(i) and (b)(i) is achieved when $\gamma \rightarrow \infty$.

Proof. See Appendix. \square

Remark 1. Part (a) of this lemma shows us how the entropy at s_0 provides an upper bound on a frequency weighted H_2 -norm of G . Perhaps more importantly, part (b) shows us how the entropy at infinity provides an upper bound on the usual H_2 -norm of G , if it exists. It is well-known that $\|G\|_2^2$, with $G = \mathcal{F}(P, K)$, is just the LQG cost associated with the plant P and controller K . So part (b)(i) of the lemma gives us a guaranteed upper bound on the LQG cost, which indicates that the minimum entropy/ H_2 problem has a combined H_2 /LQG interpretation. In fact, this connection is quite deep and in Mustafa (1989a) an equivalence with the combined

H_2 /LOG approach of Bernstein and Haddad (1989) is proved

Remark 2. Lemma 1 also shows that relaxing the H_2 -norm constraint entirely by allowing $\gamma \rightarrow \infty$ gives equality in both parts (a)(i) and (b)(i). In particular, part (b)(i) becomes $I(G; \infty; \infty) = \|G\|_2^2$. Minimizing the entropy at infinity in this case is therefore equivalent to minimizing $\|G\|_2^2$, i.e. the H_2 -optimal (or LOG) problem is recovered. For brevity, we shall not discuss the full implications here. A fuller discussion of some of the general properties of minimum entropy H_2 controllers and the connections with LOG control, together with some numerical results, may be found in Mustafa (1989b).

Remark 3. The risk-sensitive LOG problem (see e.g. Whittle, 1981) involves minimizing an exponential-of-quadratic cost. It may be viewed as a generalization of the usual LOG problem. It was shown in Glover and Doyle (1988) that the minimum entropy/ H_2 problem (at infinity) considered here is equivalent to the infinite-time risk-sensitive LOG problem. This result provides another interesting interpretation of the entropy problem, and also prompted Whittle (1989) to give a direct derivation of the entropy minimizing property of the optimal LEOG controller.

Remark 4. The minimum entropy control problem studied in this paper is related to signal processing via the problem of finding the positive definite "band extension" of a given operator. To see the connection, use a well-known fact (Theorem 7.7.6 of Horn and Johnson, 1985) and the definition of the H_2 -norm to show that

$$\left\| \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^* & R_{22} + Q \end{bmatrix} \right\|_2 = \gamma$$

if and only if

$$M = \begin{bmatrix} I & 0 & \gamma^{-1}R_{12} & \gamma^{-1}(R_{12} + Q) \\ 0 & I & \gamma^{-1}R_{11} & \gamma^{-1}R_{12} \\ \gamma^{-1}R_{12}^* & \gamma^{-1}R_{11}^* & I & 0 \\ \gamma^{-1}(R_{12} + Q)^* & \gamma^{-1}R_{12}^* & 0 & I \end{bmatrix} \quad (6) \quad \forall \gamma \in \mu_0 \cap \mathbb{R}^+$$

Thus we seek a positive definite extension M of the "band" data

$$N := \begin{bmatrix} I & 0 & \gamma^{-1}R_{12} & \gamma^{-1}R_{12} \\ 0 & I & \gamma^{-1}R_{11} & \gamma^{-1}R_{12} \\ \gamma^{-1}R_{12}^* & \gamma^{-1}R_{11}^* & I & 0 \\ \gamma^{-1}R_{12}^* & \gamma^{-1}R_{11}^* & 0 & I \end{bmatrix}$$

This can be interpreted as "band" data because only the anticausal component on R_{22} is specified. In Dym and Gohberg (1986, 1988) the *band extension* of N is defined as $M \geq 0$ in (7) such that M^{-1} has the same banded structure as N . It is shown in Dym and Gohberg (1986, 1988) that this unique band extension also minimizes the entropy. Although the very general results of Dym and Gohberg (1986, 1988) could be applied to our problem, we choose to adapt the method of Arov and Krein (1983). This makes for a relatively short and self-contained derivation.

3. Derivation of the minimum entropy solution

In this section we solve the following minimum entropy distance problem, as posed earlier. Let $y_0 \in \mathbb{R}^+$ and let

$$R = \begin{matrix} & \begin{matrix} m_1 & p_1 \end{matrix} \\ \begin{matrix} p_1 & m_1 \end{matrix} & \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \end{matrix}$$

be given where

$$R^* \in \mathcal{RH}_\infty, \quad m_1 \leq p_2, \quad p_1 \leq m_2 \quad (8)$$

Define the error system E by

$$E = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q \end{bmatrix}, \quad Q \in \mathcal{RH}_\infty$$

and let

$$\gamma_{\text{opt}} = \inf \{ \|E\|_2 : Q \in \mathcal{RH}_\infty \}$$

Then for $\gamma \geq \gamma_{\text{opt}}$ find $\hat{Q} \in \mathcal{RH}_\infty$ such that the entropy $I(E, \gamma, y_0)$ is minimized over those E which satisfy $\|E\|_2 = \gamma$.

We have seen that this problem is equivalent to finding a stabilizing controller K which keeps $\| \Phi(P, K) \|_2 \leq \gamma$ and minimizes the closed-loop entropy.

Solution proceeds by firstly parametrizing all E which satisfy the bound $\|E\|_2 \leq \gamma$. Such a parametrization is given in Ball and Cohen (1987), in terms of a linear fractional map of a J -unitary matrix and an arbitrary, stable contraction Φ (that is, $\Phi \in \mathcal{RH}_\infty$), but it is more convenient to use the parametrization of Glover *et al.* (1990) in terms of the linear fractional map of an all-pass matrix and an arbitrary, stable contraction Φ . By adapting the method of Arov and Krein (1983), we are able to derive the unique choice of Φ which minimizes the entropy and a value for the minimum entropy, for both the general case of $y_0 \in \mathbb{R}^+$ and when $y_0 = \infty$.

3.1 The general case Here we solve the minimum entropy distance problem for arbitrary $y_0 \in \mathbb{R}^+$ and E proper, but not necessarily strictly proper. The class of error systems E over which the entropy must be minimized is parametrized in the following lemma.

Lemma 2 [Glover *et al.* (1990)] All solutions

$$E = \begin{matrix} & \begin{matrix} m_1 & p_1 \end{matrix} \\ \begin{matrix} p_1 & m_1 \end{matrix} & \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q \end{bmatrix} \end{matrix}$$

with

$$R^* \hat{Q} \in \mathcal{RH}_\infty, \quad m_1 \leq p_2, \quad p_1 \leq m_2$$

to the distance problem $\|E\|_2 = \gamma$, where $\gamma \geq \gamma_{\text{opt}}$, are given by

$$E = \gamma \Phi (R_{\text{ap}} + Q_{\text{ap}} \Psi),$$

where

$$\Phi = \begin{matrix} & \begin{matrix} m_1 & p_1 \end{matrix} \\ \begin{matrix} p_1 & m_1 \end{matrix} & \begin{bmatrix} 0 & 0 \\ 0 & \Phi \end{bmatrix} \end{matrix}, \quad \Phi \in \mathcal{RH}_\infty \quad (9)$$

Also,

$$R_{\text{ap}} + Q_{\text{ap}} = \begin{matrix} & \begin{matrix} m_1 & p_1 \end{matrix} \\ \begin{matrix} p_1 & m_1 \end{matrix} & \begin{bmatrix} [R_{\text{ap}} + Q_{\text{ap}}]_{11} & [R_{\text{ap}} + Q_{\text{ap}}]_{12} \\ [R_{\text{ap}} + Q_{\text{ap}}]_{21} & [R_{\text{ap}} + Q_{\text{ap}}]_{22} \end{bmatrix} \end{matrix}$$

$$\Psi = \begin{matrix} & \begin{matrix} m_1 & p_1 \end{matrix} & \begin{matrix} p_1 & m_1 \end{matrix} & \begin{matrix} m_2 & p_2 \end{matrix} \\ \begin{matrix} p_1 & m_1 \end{matrix} & \begin{bmatrix} \gamma^{-1}R_{11} & \gamma^{-1}R_{12} & R_{13} & 0 \\ \gamma^{-1}R_{21} & \gamma^{-1}(R_{22} + Q_{22}) & R_{23} + Q_{23} & Q_{24} \\ R_{31} & R_{32} + Q_{32} & R_{33} + Q_{33} & Q_{34} \\ 0 & Q_{42} & Q_{43} & Q_{44} \end{bmatrix} \end{matrix}$$

Further, $R_{\text{ap}}^* = Q_{\text{ap}} \in \mathcal{RH}_\infty$, $R_{\text{ap}} + Q_{\text{ap}}$ is all-pass and $Q_{44}(z) = 0$.

State-space realizations of R_{ap} and Q_{ap} are available in Glover *et al.* (1990), in terms of the realization of R and the solutions to two algebraic Riccati equations. These realizations will be stated and used in Section 3.2.

The next lemma relates the entropy of the linear fractional map of an all-pass matrix J and an arbitrary stable contraction Ψ to the entropy of Ψ itself.

Lemma 3. Suppose

$$J = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix}$$

is all-pass, $\Psi \in \mathcal{RH}_\infty$ and $J_{22}\Psi \in \mathcal{RH}_\infty$. Then

$$I(\gamma\mathcal{F}(J, \Psi); \gamma, s_0) = \gamma^2 I(\Psi; 1, s_0) + \gamma^2 I(J_{11}; 1, s_0) \\ + \gamma^2 (\operatorname{Re} s_0) \ln |\det (I - J_{22}(s_0)\Psi(s_0))|.$$

Proof Throughout this proof take $s = j\omega$. Since J_{22} is part of an all-pass matrix, $\|J_{22}\|_\infty \leq 1$. Therefore $\|J_{22}\Psi\|_\infty < 1$ because $\|\Psi\|_\infty < 1$. This, together with the assumption that $J_{22}\Psi \in \mathcal{RH}_\infty$, implies that $\det (I - J_{22}\Psi)$ is a unit in \mathcal{RH}_∞ . As J is all-pass, we then have that (Redheffer, 1960)

$$[\gamma\mathcal{F}(J, \Psi)]^*[\gamma\mathcal{F}(J, \Psi)] = \gamma^2 I,$$

so the entropy $I(\gamma\mathcal{F}(J, \Psi); \gamma, s_0)$ is well-defined. Using

$$J^*(j\omega)J(j\omega) = I \quad (10)$$

in block-partitioned form, it is straightforward to show that

$$I = \gamma^{-2} [\gamma\mathcal{F}(J, \Psi)]^*[\gamma\mathcal{F}(J, \Psi)] \\ = J_{21}^*[I - J_{22}\Psi]^* [I - \Psi^*\Psi][I - J_{22}\Psi]^{-1} J_{21}.$$

From this, and the fact that for any square real-rational transfer function matrix G

$$\ln |\det (G^*G)| = \ln |\det (G)| + \ln |\det (G^*)| = 2 \ln |\det (G)|,$$

we obtain

$$\ln |\det (I - \gamma^{-2} [\gamma\mathcal{F}(J, \Psi)]^*[\gamma\mathcal{F}(J, \Psi)])| \\ = \ln |\det (I - \Psi^*\Psi)| \\ + \ln |\det (J_{21}^* J_{21})| = 2 \ln |\det (I - J_{22}\Psi)|.$$

Substituting this into the definition of $I(\gamma\mathcal{F}(J, \Psi); \gamma, s_0)$ and using the (1,1) block of equation (10) to write $J_{21}^* J_{21} = I - J_{11}^* J_{11}$ it follows that

$$I(\gamma\mathcal{F}(J, \Psi); \gamma, s_0) = \gamma^2 I(\Psi; 1, s_0) + \gamma^2 I(J_{11}; 1, s_0) \\ + \gamma^2 \int_{-\pi}^{\pi} \ln |\det (I - J_{22}(j\omega)\Psi(j\omega))| \frac{\operatorname{Re} s_0}{|s_0 - j\omega|} d\omega. \quad (11)$$

But, from above, $\det (I - J_{22}\Psi)$ is a unit in \mathcal{RH}_∞ , which permits the use of Poisson's Integral Theorem [see e.g. p. 343 of Rudin (1986); as done in Limebeer and Hung (1987)] to evaluate the integral in (11), giving

$$I(\gamma\mathcal{F}(J, \Psi); \gamma, s_0) = \gamma^2 I(\Psi; 1, s_0) + \gamma^2 I(J_{11}; 1, s_0) \\ + \gamma^2 (\operatorname{Re} s_0) \ln |\det (I - J_{22}(s_0)\Psi(s_0))|$$

as claimed. \square

We are now in a position to derive the unique, stable, contractive Φ in the parametrization of all error systems, which minimizes the entropy $I(E; \gamma, s_0)$.

Theorem 1. Consider the class of error systems E which satisfy the condition $\|E\|_\infty \leq \gamma$ as parametrized in Lemma 2 by

$$E = \gamma\mathcal{F}\left(R_{aa} + Q_{aa}, \begin{bmatrix} 0 & 0 \\ 0 & \Phi \end{bmatrix}\right), \quad \Phi \in \mathcal{RH}_\infty \quad (12)$$

Then the entropy $I(E; \gamma, s_0)$ is minimized over this class of E by the unique choice

$$\Phi = Q_{aa}^*(s_0).$$

Proof Here we adapt the approach of Arov and Krein (1983), also used in Limebeer and Hung (1987), to the present setting. Lemma 2 gives all error systems in the form (12), where $R_{aa} + Q_{aa}$ is all-pass. Also,

$$[R_{aa} + Q_{aa}]_{22} \begin{bmatrix} 0 & 0 \\ 0 & \Phi \end{bmatrix} = \begin{bmatrix} 0 & Q_{aa}\Phi \\ 0 & Q_{aa}\Phi \end{bmatrix} \in \mathcal{RH}_\infty,$$

because Q_{aa} , Q_{aa} and Φ are all in \mathcal{RH}_∞ . Hence, we may apply Lemma 3 to E to obtain

$$I(E; \gamma, s_0) = \gamma^2 I(\Phi; 1, s_0) + \gamma^2 I([R_{aa} + Q_{aa}]_{11}; 1, s_0) \\ + \gamma^2 (\operatorname{Re} s_0) \ln |\det (I - Q_{aa}(s_0)\Phi(s_0))| \quad (13)$$

If $Q_{aa}(s_0) = 0$ then

$$I(E; \gamma, s_0) = \gamma^2 I(\Phi; 1, s_0) + \gamma^2 I([R_{aa} + Q_{aa}]_{11}; 1, s_0),$$

which is clearly minimized by the unique choice $\Phi = 0 = Q_{aa}^*(s_0)$, and there is nothing more to prove. So, henceforth in this proof assume $Q_{aa}(s_0) \neq 0$. Define the constant Julia matrix H (see e.g. p. 148 of Young, 1988) by

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \\ = \begin{bmatrix} -Q_{aa}^*(s_0) & (I - Q_{aa}^*(s_0)Q_{aa}(s_0))^{1/2} \\ (I - Q_{aa}(s_0)Q_{aa}^*(s_0))^{1/2} & Q_{aa}(s_0) \end{bmatrix}$$

where $(\cdot)^{1/2}$ denotes Hermitian square root. It is easy to verify that H is unitary. Also,

$$H_{22}\Phi(s) = Q_{aa}(s_0)\Phi(s)$$

which is in \mathcal{RH}_∞ . Let us map the unit ball in \mathcal{RH}_∞ onto itself by the linear fractional map

$$\tilde{\Phi} := \mathcal{F}(H, \Phi), \quad \Phi \in \mathcal{RH}_\infty.$$

Note that this maps $\Phi = Q_{aa}^*(s_0)$ onto $\tilde{\Phi} = 0$. Lemma 3 is applicable:

$$I(\tilde{\Phi}; 1, s_0) = I(\Phi; 1, s_0) + I(Q_{aa}^*(s_0); 1, s_0) \\ + (\operatorname{Re} s_0) \ln |\det (I - Q_{aa}(s_0)\Phi(s_0))|$$

Use this together with (13) to relate the entropy of E to the entropy of $\tilde{\Phi}$:

$$I(E; \gamma, s_0) = \gamma^2 I(\tilde{\Phi}; 1, s_0) + \gamma^2 I([R_{aa} + Q_{aa}]_{11}; 1, s_0) \\ - \gamma^2 I(Q_{aa}^*(s_0); 1, s_0), \quad (14)$$

from which it is immediate that $I(E; \gamma, s_0)$ is minimized by the unique choice $\tilde{\Phi} = 0$. But from above, $\tilde{\Phi} = 0 \Leftrightarrow \Phi = Q_{aa}^*(s_0)$, and the theorem is proved. \square

Denote minimum entropy quantities by $(\cdot)_{\min}$. An expression for the minimum value of the entropy follows with ease from the above proof.

Corollary 1.

$$I(E_{\min}; \gamma, s_0) = \gamma^2 I([R_{aa} + Q_{aa}]_{11}; 1, s_0) \\ - \gamma^2 I(Q_{aa}^*(s_0); 1, s_0) \quad (15) \\ = \gamma^2 (\operatorname{Re} s_0) [-\ln |\det (R_{aa}^*(s_0))| \\ - \ln |\det (Q_{aa}(s_0))| \\ + (1/2) \ln |\det (I - Q_{aa}^*(s_0)Q_{aa}(s_0))|] \quad (16)$$

Proof Equation (15) follows immediately from equation (14) on setting $\tilde{\Phi} = 0$. To show (16), recall that $R_{aa} + Q_{aa}$ is all-pass i.e.

$$(R_{aa} + Q_{aa})^*(R_{aa} + Q_{aa}) = I, \quad \forall s = j\omega$$

The (1,1) block of this gives, $\forall s = j\omega$,

$$I - [R_{aa} + Q_{aa}]_{11}^*[R_{aa} + Q_{aa}]_{11} = [R_{aa} + Q_{aa}]_{21}^*[R_{aa} + Q_{aa}]_{21}$$

so that, along the imaginary axis,

$$\ln |\det (I - [R_{aa} + Q_{aa}]_{11}^*[R_{aa} + Q_{aa}]_{11})| \\ = 2 \ln |\det ([R_{aa} + Q_{aa}]_{21})| \quad (17) \\ = 2 \ln |\det (R_{aa}^*)| + 2 \ln |\det (Q_{aa})|, \quad (18)$$

where (18) follows from (17) on examination of the structure of $R_{aa} + Q_{aa}$ in Lemma 3.1. Substituting (18) into the definition of entropy, we see that

$$\gamma^2 I([R_{aa} + Q_{aa}]_{11}; 1, s_0) = -\frac{\gamma^2}{\pi} \int_{-\pi}^{\pi} \{\ln |\det (R_{aa}^*(j\omega))| \\ + \ln |\det (Q_{aa}(j\omega))|\} \left[\frac{\operatorname{Re} s_0}{|s_0 - j\omega|} \right] d\omega. \quad (19)$$

Since R_{aa}^* and Q_{aa} are units in \mathcal{RH}_∞ (Glover *et al.*, 1990),

Poisson's Integral Theorem may be used to evaluate (19) as

$$\gamma^2 I([R_{\infty} + Q_{\infty}]_{11}; 1; s_0) = -\gamma^2 (\operatorname{Re} s_0) \times (\ln |\det(R_{\infty}^*(s_0))| + \ln |\det(Q_{\infty}(s_0))|) \quad (20)$$

The second term in (15) is

$$\begin{aligned} & -\gamma^2 I(Q_{\infty}^*(s_0); 1; s_0) \\ &= \frac{\gamma^2}{2\pi} \ln |\det(I - Q_{\infty}(s_0)Q_{\infty}^*(s_0))| \int_{\pi}^{\infty} \left[\frac{\operatorname{Re} s_0}{|s_0 - j\omega|} \right]^2 d\omega \\ &= \frac{\gamma^2}{2\pi} \ln |\det(I - Q_{\infty}(s_0)Q_{\infty}^*(s_0))| \cdot \pi \cdot (\operatorname{Re} s_0) \end{aligned}$$

and this with (20) gives (16). \square

3.2. Entropy at infinity. We turn our attention in this section to the special case of entropy at infinity i.e. when $s_0 \rightarrow \infty$ along the real axis. See Dym (1989) for entropy at infinity in a different setting, and Gohberg *et al.* (1988). As remarked in Section 2, this is the most interesting and important case because of the strong connections with other control problems.

The entropy at infinity of G , (where $\|G\|_{\infty} < \gamma$) is

$$I(G; \gamma; \infty) = -\frac{\gamma^2}{2\pi} \int_{\pi}^{\infty} \ln |\det(I - \gamma^{-2} G^*(j\omega)G(j\omega))| d\omega$$

which is finite if G is strictly proper. For our problem, where we minimize the entropy of the error system E , this means that the minimum value of the entropy at infinity is finite if E_{∞} is strictly proper; this occurs when R in the distance problem of Lemma 2 is strictly proper. This in turn occurs when $P_{11}(\infty)$ is strictly proper. Referring to the standard configuration of Fig. 1, this corresponds to no direct feedthrough terms from the exogenous inputs w to controlled outputs z .

The results for the minimum entropy problem at infinity are particularly simple. The minimum entropy solution is obtained by setting the arbitrary stable contraction Φ to zero i.e. by choosing the "central solution" out of the set of admissible E , and an explicit formula for the minimum value of the entropy is derived in terms of the state-space realizations inherent in the solution of the distance problem of Lemma 2; these state-space realizations are stated in the next lemma.

Lemma 4 (Glover *et al.* 1990). Consider the distance problem of Lemma 2 in the case $R(\infty) = 0$. Suppose R has a realization

$$R = \begin{array}{c} \begin{array}{c} z_n \uparrow \\ \mu_1 \downarrow \\ \mu_2 \downarrow \end{array} \begin{array}{c} \xrightarrow{m_1} \xrightarrow{p_1} \xrightarrow{p_2} \\ \left[\begin{array}{c|c|c} A & B_1 & B_2 \\ \hline C_1 & 0 & 0 \\ \hline C_2 & 0 & 0 \end{array} \right] \end{array} \end{array}$$

Then $R_{\infty} + Q_{\infty}$, as in the parametrization of all solutions to the distance problem $\|E\|_{\infty} < \gamma$ given in Lemma 2, has a realization

$$R_{\infty} + Q_{\infty} = \left[\begin{array}{c|c} \tilde{A} & \tilde{E} \\ \hline \tilde{C} & \tilde{D} \end{array} \right]$$

where

$$\begin{aligned} \tilde{A} &= \begin{bmatrix} A & 0 \\ 0 & \tilde{A} \end{bmatrix} \\ \tilde{B} &= \begin{bmatrix} \gamma^{-1/2} B_1 & \gamma^{-1/2} B_2 & -\gamma^{-1/2} X C_1^T & 0 \\ 0 & \gamma^{-1/2} Y Z^{-1} B_2 & -\gamma^{-1/2} C_1^T & \gamma^{-1/2} Z^{-1} C_2^T \end{bmatrix} \\ \tilde{C} &= \begin{bmatrix} \gamma^{-1/2} C_1 & 0 \\ \gamma^{-1/2} C_2 & -\gamma^{-1/2} C_2^T X \\ -\gamma^{-1/2} B_1^T Y & \gamma^{-1/2} B_1^T Z^{-1} \\ 0 & -\gamma^{-1/2} B_2^T \end{bmatrix} \\ \tilde{D} &= \begin{bmatrix} 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{bmatrix} \end{aligned}$$

and where

$$X = \operatorname{Ric} \left[\begin{array}{c|c} A^T & \gamma^{-1/2} C_1^T C_1 \\ \hline -B_1 B_1^T & B_2 B_2^T \end{array} \right] \quad (21)$$

$$Y = \operatorname{Ric} \left[\begin{array}{c|c} A & \gamma^{-1/2} B_1 B_1^T \\ \hline -C_1^T C_1 & -C_2^T C_2 \end{array} \right] \quad (22)$$

$$Z = \gamma^{-1/2} X Y^{-1}$$

$$\tilde{A} = A^T - \gamma^{-1/2} Y Z^{-1} B_2 B_2^T - \gamma^{-1/2} C_2^T C_1 X$$

Applying the results of the previous section using this realization, and taking $s_0 \rightarrow \infty$, gives us the following important theorem.

Theorem 2. The entropy at infinity $I(E; \gamma; \infty)$, is minimized over the class of error systems E in the distance problem of Lemma 2, by the unique choice $\Phi = 0$.

If $R(\infty) = 0$, then the minimum entropy error system is simply

$$\begin{aligned} E_{\infty} &= \gamma [R_{\infty} + Q_{\infty}]_{11} \\ &= \left[\begin{array}{c|c|c} A & 0 & B_1 & B_2 \\ \hline 0 & A & 0 & \gamma^{-1/2} Y Z^{-1} B_2 \\ \hline C_1 & 0 & 0 & 0 \\ \hline C_2 & \gamma^{-1/2} C_2 X & 0 & 0 \end{array} \right] \quad (23) \end{aligned}$$

and the minimum value of the entropy is

$$I(E_{\infty}; \gamma; \infty) = -\operatorname{trace}[B_1^T Y B_1] + \operatorname{trace}[B_2^T Y Z^{-1} B_2] \quad (24)$$

$$= -\operatorname{trace}[C_1 X C_1^T] + \operatorname{trace}[C_2 Z^{-1} X C_2^T] \quad (25)$$

Proof. From Theorem 1, the minimum entropy error system is characterized by the unique choice $\Phi = Q_{\infty}^*(s_0)$. Letting $s_0 \rightarrow \infty$ along the real axis gives the minimum entropy at infinity choice as $\Phi = Q_{\infty}^*(\infty) = 0$, because Q_{∞} is strictly proper from Lemma 2. For details of the limiting argument see Theorem 3.1 of Glover and Mustafa (1989). That the minimum entropy error system has a realization (23) follows easily by setting $\Phi = 0$ in (9) and using Lemma 4.

To obtain the minimum value of the entropy, we take the limit as $s_0 \rightarrow \infty$ along the real axis of the result of Corollary 1. That is,

$$\begin{aligned} I(E_{\infty}; \gamma; \infty) &= \lim_{s_0 \rightarrow \infty} (\gamma \ln |\det(R_{\infty}^*(s_0))| \\ &\quad - \ln |\det(Q_{\infty}(s_0))| \\ &\quad + (1/2) \ln |\det(I - Q_{\infty}^*(s_0)Q_{\infty}(s_0))|) \quad (26) \end{aligned}$$

Consider a typical term from (26)

$$\lim_{s_0 \rightarrow \infty} (s_0 \ln |\det(I + C(s_0 I - A)^{-1} B)|)$$

Note we have dropped the modulus sign because s_0 is real here. Now,

$$C(s_0 I - A)^{-1} B = C B s_0^{-1} + O(s_0^{-2}),$$

so by Lemma A.1 (of the Appendix) we have

$$s_0 \ln |\det(I + C(s_0 I - A)^{-1} B)| = s_0 (\operatorname{trace}[C B s_0^{-1}] + O(s_0^{-2}))$$

Therefore, on taking limits as $s_0 \rightarrow \infty$

$$\lim_{s_0 \rightarrow \infty} (s_0 \ln |\det(I + C(s_0 I - A)^{-1} B)|) = \operatorname{trace}[C B]$$

A similar result for scalar systems has been derived independently in a different context in Anderson and Mungo (1985). Apply this to the terms in (26) using

$$R_{\infty}^*(s_0) = I + \gamma^{-1/2} B_1^T (s_0 I + A^T)^{-1} Y B_1$$

$$Q_{\infty}(s_0) = I + \gamma^{-1/2} B_2^T (s_0 I - \tilde{A})^{-1} Y Z^{-1} B_2$$

from Lemma 2, to get

$$I(E_{\infty}; \gamma; \infty) = -\operatorname{trace}[B_1^T Y B_1] + \operatorname{trace}[B_2^T Y Z^{-1} B_2]$$

as required. Note that the third term in (26) is zero in the limit because Q_{∞} is strictly proper.

The dual expression (25) follows in an entirely similar fashion; one notes that $I(E_{\text{MI}}; \gamma; x) = I(E_{\text{MI}}^T; \gamma; x)$, leading to

$$I(E_{\text{MI}}; \gamma; x) = \lim_{s_0 \rightarrow \infty} (\gamma^2 s_0 \{-\ln |\det(R_{11}^*(s_0))| \\ - \ln |\det(Q_{22}(s_0))| \\ + (1/2) \ln |\det(I - Q_{22}(s_0)Q_{22}^*(s_0))|\}),$$

which gives (25) in the limit. \square

Remark 5. Notice that the entropy formulae (24) and (25) depend only on the state-space realization of R and the solutions X and Y to the two Riccati equations (21) and (22) which are inherent in the solution to the distance problem. Calculation of the minimum value of the entropy therefore imposes negligible extra computational problems. Furthermore, the minimum entropy error system (23), being the linear fractional map of $\Phi = 0$, is simply γ times the p_1 by m_1 (1, 1) block of $R_{\text{ss}} + Q_{\text{ss}}$, which is also available from the solution to the distance problem with no extra computation.

Remark 6. Recall, from Lemma 1 and Remark 2, that $I(G; x; x) = \|G\|_2^2$ for strictly proper G . Thus if we let $\gamma \rightarrow \infty$ in our minimum entropy solution we should obtain exactly the H_2 -optimal solution. We show here that this is indeed the case.

By using Wimmer (1985), it may be shown that the positive semidefinite matrices $-X_{\gamma \rightarrow \infty}$ and $-Y_{\gamma \rightarrow \infty}$ are monotonically decreasing as γ increases. Taking $\gamma \rightarrow \infty$ we obtain

$$X_{\gamma \rightarrow \infty} = \text{Ric} \begin{pmatrix} 0 \\ -B_1 B_1^T - B_2 B_2^T \\ -A \end{pmatrix} \quad (27)$$

$$Y_{\gamma \rightarrow \infty} = \text{Ric} \begin{pmatrix} 0 \\ -C_1^T C_1 - C_2^T C_2 \\ -A^T \end{pmatrix} \quad (28)$$

and

$$Z_{\gamma \rightarrow \infty} = -I,$$

(with an obvious notation). Equations (27) and (28) identify the matrices $-X_{\gamma \rightarrow \infty}$ and $-Y_{\gamma \rightarrow \infty}$ as the controllability and observability Gramians of $R(\cdot, \infty)$, respectively. Using this fact, a simple calculation shows that

$$I(E_{\text{MI}}; \infty; x) = \text{trace} \{B_1^T Y_{\gamma \rightarrow \infty} B_1\} \\ + \text{trace} \{B_2^T Y_{\gamma \rightarrow \infty} Z_{\gamma \rightarrow \infty} B_2\} \\ = \text{trace} \{[B_1 \ B_2]^T [-Y_{\gamma \rightarrow \infty}] [B_1 \ B_2]\} \\ = \|R\|_2^2. \quad (29)$$

Also, inspection of (23) as $\gamma \rightarrow \infty$ leads to $\dot{Q}_{\text{MI}} = 0$. It is well-known that the $Q \in \mathcal{RH}$, which minimizes

$$\|E\|_2 = \begin{vmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{vmatrix} \dot{Q}$$

is $\dot{Q} = Q_{H_2} = 0$ (the H_2 -optimal solution) and in that case $\|E_{H_2}\|_2 = \|R\|_2$. Comparing this with (29) shows that we have $I(E_{\text{MI}}; \infty; x) = \|E_{H_2}\|_2^2$, illustrating the equivalence between the minimum entropy distance problem at $\gamma, s_0 \rightarrow \infty$ and the H_2 -optimal distance problem.

Remark 7. Note that if $\gamma \rightarrow \gamma_{\text{opt}}$, then $E_{\text{MI}} \rightarrow E_{\text{opt}}$, the H_2 -optimal solution. Also (Remark 6), if $\gamma \rightarrow \infty$, then $E_{\text{MI}} \rightarrow E_{H_2}$. It is easily shown that $I(E_{\text{MI}}; \gamma; x)$ is a monotonically decreasing function of γ . Thus γ can be used to move from H_2 -optimal to H_∞ -optimal via the minimum entropy solutions for $\gamma_{\text{opt}} \leq \gamma < \infty$.

4. Conclusion

We have posed, discussed and solved the problem of minimizing the entropy of a general H_∞ control system. The solution was obtained by solving the minimum entropy version of the equivalent distance problem—the distance problem being, until recently, a necessary stepping stone to the solution of most H_∞ control problems. In the case of greatest interest, when entropy is evaluated at infinity, the minimum entropy solution was shown to be the central member of the admissible class and to be an interesting

compromise between H_∞ and LQG or H_2 control. Beginning with the relevant state-space data for the distance problem, explicit state-space formulae for this solution, together with the value of its entropy, were derived in terms of just two algebraic Riccati equations. Since the relevant state-space data for the distance problem are determined precisely by the standard plant in the original minimum entropy H_∞ control problem, back-substitution could be carried out to find the minimum entropy controller in terms of the standard plant only. It is, in fact, possible to bypass the distance problem entirely and solve the minimum entropy H_∞ control problem directly in terms of the given standard plant, as is done in Glover and Mustafa (1989).

References

- Anderson, B. D. O. and D. L. Mingori (1985). Use of frequency dependence in Linear Quadratic control problems to frequency-shape robustness. *J. Guidance Control Dynam.*, **8**, 397–401.
- Arov, D. and M. G. Krein (1981). Problem of search of the minimum entropy in indeterminate extension problems. *Funct. Anal. Applic.*, **15**, 123–126.
- Arov, D. and M. G. Krein (1983). On the evaluation of entropy functionals and their minima in generalized extension problems. *Acta Scientia Math.*, **45**, 33–50 (in Russian).
- Ball, J. A. and N. Cohen (1987). Sensitivity minimization in H_∞ -norm. Parametrization of all suboptimal solutions. *Int. J. Control*, **46**, 785–816.
- Bernstein, D. S. and W. M. Haddad (1989). LQG control with an H_∞ performance bound. A Riccati equation approach. *IEEE Trans. Aut. Control*, **AC-34**, 293–305.
- Doyle, J. C., K. Glover, P. P. Khargonekar and B. A. Francis (1989). State-space solutions to standard H_2 and H_∞ control problems. *IEEE Trans. Aut. Control*, **AC-34**, 831–849.
- Dym, H. (1989). *J-Contractive Matrix Functions, Reproducing Kernel Hilbert Spaces and Interpolation*. Volume 71 of the *Regional Conference Series in Mathematics*. American Mathematical Society.
- Dym, H. and I. Gohberg (1986). A maximum entropy principle for contractive interpolants. *J. Funct. Anal.*, **65**, 83–125.
- Dym, H. and I. Gohberg (1988). A new class of contractive interpolants and maximum entropy principles. In I. Gohberg (Ed.), *Topics in Operator Theory and Interpolation. Operator Theory: Advances and Applications*, **29**, 117–150. Birkhäuser, Stuttgart.
- Francis, B. A. (1987). *A Course in H_∞ Control Theory*. Volume 88 of *Lecture Notes in Control and Information Sciences*. Springer, Berlin.
- Gantmacher, F. R. (1959). *The Theory of Matrices*. Chelsea, New York.
- Glover, K. and J. C. Doyle (1988). State-space formulae for all stabilizing controllers that satisfy an H_∞ -norm bound and relations to risk sensitivity. *Syst. Control Lett.*, **11**, 167–172.
- Glover, K., D. J. N. Limebeer, J. C. Doyle, E. M. Kasenally and M. G. Safonov (1990). A characterization of all solutions to the four block general distance problem (to appear). *SIAM J. Control Optimiz.*
- Glover, K. and D. Mustafa (1989). Derivation of the maximum entropy H_∞ -controller and a state-space formula for its entropy. *Int. J. Control*, **50**, 899–916.
- Gohberg, I., M. A. Kaashoek and F. van Schagen (1988). Rational contractive and unitary interpolants in realized form. *Integral Equations and Operator Theory*, **11**, 105–127.
- Horn, R. A. and C. R. Johnson (1985). *Matrix Analysis*. Cambridge University Press, Cambridge, U.K.
- Kučera, V. (1979). *Discrete Linear Control: The Polynomial Equation Approach*. Wiley, Chichester, U.K.
- Limebeer, D. J. N. and Y. S. Hung (1987). An analysis of the pole-zero cancellations in H_∞ -optimal control problems of the first kind. *SIAM J. Control Optimiz.*, **25**, 1457–1493.

Mustafa, D. (1989a). Relations between maximum entropy/ H_∞ control and combined H_∞ /LOG control. *Syst. Control Lett.*, **12**, 193–203.

Mustafa, D. (1989b). On H_∞ control, LOG control and minimum entropy. *Proc. Int. Symp. on the Mathematical Theory of Networks and Systems*. Amsterdam, (to appear). Birkhäuser, Basel.

Redheffer, R. M. (1960). On a certain linear fractional transformation. *J. Math. Physics*, **39**, 269–286.

Rudin, W. (1986). *Real and Complex Analysis*, 3rd edn. McGraw-Hill, New York.

Whittle, P. (1981). Risk-sensitive linear/quadratic/Gaussian control. *Advances in Applied Probability*, **13**, 764–777.

Whittle, P. (1989). Entropy-minimizing and risk-sensitive control rules. *Syst. Control Lett.*, **13**, 1–7.

Wimmer, H. K. (1985). Monotonicity of maximal solutions of algebraic Riccati equations. *Syst. Control Lett.*, **5**, 317–319.

Youla, A. D. C., H. A. Jabr and J. J. Bongiorno (1976). Modern Wiener-Hopf design of optimal controllers. Part II—the multivariable case. *IEEE Trans. Aut. Control*, **AC-21**, 319–338.

Young, N. (1988). *An Introduction to Hilbert Space*. Cambridge University Press, Cambridge, U.K.

Appendix A: Proof of Lemma 1

The following technical lemma is firstly needed

Lemma A.1. Let M be a real square matrix, let N be a matrix and let ϵ be a real number. Then

- (i) $-\ln \det(I - \epsilon M) = \epsilon \operatorname{trace}[M] + O(\epsilon^2)$
- (ii) $-\ln \det(I - \epsilon^2 N^* N) \leq \epsilon^2 \operatorname{trace}[N^* N]$.

Proof. Part (i). Use the Faddeev formula (see p. 88 of Vol. 1 of Gantmacher, 1959) to obtain

$$\det(I - \epsilon M) = 1 - \epsilon \operatorname{trace}[M] + O(\epsilon^2)$$

and expand the logarithm of this as a power series

Part (ii): Using the well-known inequality that $-\ln(1 -$

$x^2) \geq x^2$ for $|x| \leq 1$, we get,

$$\begin{aligned} -\ln \det(I - \epsilon^2 N^* N) &= -\sum_i \ln(1 - \epsilon^2 \lambda_i(N^* N)) \\ &\geq \sum_i \epsilon^2 \lambda_i(N^* N) \\ &= \epsilon^2 \operatorname{trace}[N^* N], \end{aligned}$$

as claimed □

We may now proceed with the proof of Lemma 1

Proof of Lemma 1. Part (a). By Lemma A.1 we can write

$$\begin{aligned} I(G, \gamma, s_0) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \operatorname{trace} \left\{ G^*(j\omega) G(j\omega) \right. \\ &\quad \left. \frac{\operatorname{Re} s_0}{(s_0 + j\omega)} d\omega + O(\gamma) \right. \\ &\quad \left. + \frac{1}{2\pi} \int_{-\infty}^{\infty} \operatorname{trace} \left\{ \left[G(j\omega) \frac{\operatorname{Re} s_0}{(s_0 + j\omega)} \right] \right. \right. \\ &\quad \left. \left. G(j\omega) \frac{\operatorname{Re} s_0}{(s_0 + j\omega)} d\omega + O(\gamma^{-2}) \right\} \right. \end{aligned}$$

Therefore,

$$I(G, \gamma, s_0) \geq \|G(s)(\operatorname{Re} s_0)/(s_0 + s)\|^2 + O(\gamma^{-2}), \quad (30)$$

which is part (a)(ii) as claimed. Noting that the $O(\gamma^{-2})$ terms are non-negative we have

$$I(G, \gamma, s_0) \geq \|G(s)(\operatorname{Re} s_0)/(s_0 + s)\|^2,$$

which is part (a)(i) as claimed, whilst

$$I(G, \gamma, s_0) \leq \|G(s)(\operatorname{Re} s_0)/(s_0 + s)\|^2$$

follows by taking $\gamma \rightarrow +\infty$ in (30).

Part (b). Firstly note that the integrands in (30) above are monotonically increasing with s_0 , and are continuous, and hence by dominated convergence both sides of (30) tend to a limit as $s_0 \rightarrow +\infty$. Each side is finite because $\|G\|_\infty < \gamma$ and $G(\infty) = 0$ by assumption. The result then follows in a similar way to the proof of part (a). □

Brief Paper

On Slowly Time-varying Systems*

MOHAMMED DAHLEH† and MUNTHER A. DAHLEH‡

Key Words—Slowly time-varying systems, frozen-time design, infinite dimensional systems

Abstract—A characterization of stabilizing controllers for slowly time-varying systems based on input-output descriptions of the plant and the controller is presented. This approach generalizes standard results on slowly time-varying systems with finite dimensional state-space representation, and allows both the plant and the controller to be infinite dimensional. The controller design will be based on frozen-time versions of the plant obtained at equally spaced instances in time.

1. Introduction

THE PROBLEM of controlling a time-varying plant arises in many applications. In the case when the plant is slowly time-varying, many control approaches have been used successfully. For example, in gain scheduling, the plant is assumed to be varying and at successive points in time a controller is designed to satisfy a set of prescribed specifications. The sequence of scheduled controllers compensate for the time-varying nature of the problem, and their performance is expected to be acceptable if the plants are varying sufficiently slowly.

The reason to study slowly time-varying systems is both theoretical and practical. The practical importance is clear from the multitude of cases documented in the literature, and the wide success of the method of gain scheduling. On the theoretical side it gives us a tool to investigate which systems can be effectively controlled by frozen-time methods, and which are the suitable design techniques. Also in a related problem, which is the problem of adaptive control, it has been shown that the study of slowly time-varying systems is crucial to the understanding of the adaptive problem especially in the presence of disturbances.

The goal of this paper is to present a new method of analyzing slowly time-varying systems which is based on an input-output approach. Most of the results in the literature are concerned with state-space approaches (Desoer and Vidyasagar, 1975), which tend to have many shortcomings. One of the major drawbacks of the state-space approach is the necessity of restricting the analysis to an *a priori* fixed-order controller, whether it is static or dynamic. This in turn makes it impossible to analyze controllers that are based on general design approaches, e.g. H^∞ , L^1 (Vidyasagar, 1985; Francis, 1987; Dahleh and Pearson, 1987a) in a unified framework. The method presented in this paper allows the consideration of infinite dimensional plants and controllers.

The results will be given for discrete-time single-input single-output plants. The plant is time-varying and the controller is designed on the basis of frozen-time versions of

the plant. A novelty of this new approach is that the controller is not necessarily adjusted for every instance of time, and thus can be used for a fixed time window before a new controller is implemented. This situation, which is the norm in applications, was not analyzed in previous research on this problem. The length of the time window now enters as another parameter in the stability analysis.

An application of the results in this paper is the study of the problem of designing L^1 optimal controllers (Dahleh and Pearson, 1987a, b) for slowly time-varying systems. It is shown that under conditions of slow variation, it makes sense to design frozen-time L^1 optimal controllers. This gives precise conditions for the feasibility of using gain-scheduled controllers to satisfy internal stability and optimal disturbance rejection. The details of this application have been reported in (Dahleh and Dahleh, 1990) and will not be discussed in this paper.

2. Preliminaries

2.1 Mathematical preliminaries This section includes preliminaries needed in the sequel. More details can be found in Desoer and Vidyasagar (1975), Hille and Phillips (1957), Vidyasagar (1985) and Willems (1971).

l^p denotes the classical Banach space on the non-negative integers, with the associated l^p norm. l^p_+ denotes the extended space of l^p .

\mathcal{L}_{l^p} denotes the algebra of linear bounded operators on l^p which are causal, i.e. for any operator $T \in \mathcal{L}_{l^p}$

$$P_k T = P_k T P_k \quad \forall k \geq 0$$

where P_k is the standard k th truncation operator. It is straightforward to show that the action of any such T on any $f \in l^p$ has the kernel representation

$$(Tf)(n) = \sum_{j=0}^n t_{nj} f(j) \quad n \geq 0$$

where $t_{nj} \in \mathbb{R}$. The induced operator norm on T is given by

$$\|T\|_{\mathcal{L}_{l^p}} = \sup_n \sum_{j=0}^n |t_{nj}|$$

With this representation, \mathcal{L}_{l^p} is identified with the set of all infinite lower triangular matrices,

$$\begin{pmatrix} t_{00} & 0 & 0 \\ t_{10} & t_{11} & 0 \\ t_{20} & t_{21} & t_{22} \\ \vdots & \vdots & \vdots \end{pmatrix}$$

with the above norm.

Let \mathcal{L}_{l^1} be the subalgebra of \mathcal{L}_{l^1} consisting of time-invariant operators. The corresponding matrix representation has a Toeplitz structure,

$$T = \begin{pmatrix} t_0 & 0 & 0 \\ t_1 & t_0 & 0 \\ t_2 & t_1 & t_0 \\ \vdots & \vdots & \vdots \end{pmatrix}$$

and the induced operator norm simplifies to

$$\|T\|_{\mathcal{L}_{l^1}} = \sum_{j=0}^{\infty} |t_j|$$

* Received 23 February 1989, revised 5 December 1989, received in final form 12 February 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor H. Kimura under the direction of Editor H. Kwakernaak.

† Department of Electrical Engineering, Texas A & M University, College Station, TX 77843, U.S.A.

‡ Department of Electrical and Computer Engineering Science and LIDS, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A. Author to whom all correspondence should be addressed.

Hence, \mathcal{L}_{11} is isometrically isomorphic to l^1 . Associated with T , is the Z-transform (Gelfand Transform) $\hat{T}(z)$

$$\hat{T}(z) = \sum_{i=0}^{\infty} t_i z^i$$

with the associated spectral norm [or H_∞ -norm (Francis, 1987; Vidyasagar, 1985)] defined as

$$\|\hat{T}\|_\infty = \sup_{|z|=1} |\hat{T}(z)|$$

It is well known (Hille and Phillips, 1957; Vidyasagar, 1985) that if $T \in \mathcal{L}_{11}$ then $T^{-1} \in \mathcal{L}_{11}$ if and only if

$$\inf_{|z|=1} |\hat{T}(z)| > 0$$

Given a set of polynomials $\hat{A}_t(z)$

$$\hat{A}_t(z) = \sum_{i=0}^t a_i(t) z^i,$$

for each fixed t , \hat{A}_t defines an operator $A_t \in \mathcal{L}_{11}$

$$(A_t f)(k) = \sum_{i=0}^k a_i(t) f(k-i) \quad \forall k \geq 0$$

On the other hand, this collection of polynomials defines an operator A on l^1_∞ defined as

$$(A f)(t) = (A_t f)(t) = \sum_{i=0}^t a_i(t) f(t-i)$$

A is precisely the lower triangular matrix

$$\begin{pmatrix} a_0(0) & 0 & 0 & \cdots \\ a_1(1) & a_0(1) & 0 & \cdots \\ a_2(2) & a_1(2) & a_0(2) & \cdots \end{pmatrix}$$

For convenience, the following notation is adopted (Dahleh and Dahleh, 1990; Goodwin and Sin, 1984),

$$A_t B_t(z) = \sum_{i=0}^t \sum_{j=0}^t a_i(t) b_j(t) z^{i+j}$$

$$\hat{A}_t \hat{B}_t(z) = \sum_{i=0}^t \sum_{j=0}^t a_i(t) b_j(t) z^{i+j}$$

The first product is the transform of the standard composition of the linear time-invariant operators A_t, B_t for some fixed t . The second product is the composition of the time-varying operators associated with each of the families of polynomials $\hat{A}_t, \hat{B}_t, t = 0, 1, 2, \dots$. Equivalently, the matrix representation of the time-varying operator associated with $\hat{A}_t, \hat{B}_t, t = 0, 1, 2, \dots$ is precisely the product of the matrix representation of each of the individual families. (It makes sense to transform every row in the matrix if z is thought of as the unit shift operator.)

The operator A associated with $\hat{A}_t, t = 0, 1, 2, \dots$, is slowly time-varying if there exists a constant γ such that,

$$A_t = A_{t+1} + \gamma |t - t| \quad \forall t, t.$$

Such operators are denoted by $\text{STV}(\gamma)$. Equivalently, $A_t \in \text{STV}(\gamma)$ if the associated operator A is slowly time-varying.

Given an operator $T \in \mathcal{L}_{11}$, the *Integral Time Absolute Error* ITAE is defined as

$$\text{ITAE}(T) = \sum_{k=0}^{\infty} k |t_k|$$

Let $\hat{T}(z)$ be the associated Z-transform of $T, \hat{T}'(z) = (d/dz)\hat{T}(z)$, and T' is the associated time-invariant operator, referred to as the *derivative operator*, then

$$\text{ITAE}(T) = \|\hat{T}'\|_{\mathcal{L}_{11}}$$

Clearly, not all operators in \mathcal{L}_{11} have bounded ITAE. Also, boundedness of the *derivative operator* in the H_∞ -norm does not guarantee boundedness in the \mathcal{L}_{11} -norm. As will be

discussed later on, conditions in terms of the H_∞ -norm are easier to check, and therefore a characterization of when an operator has a bounded ITAE in terms of an H_∞ condition can be inferred from Section 5.

2.2. Process model A general process model can be described as

$$y = A^{-1} B u$$

where $A, B \in \mathcal{L}_{1\infty}$ and A^{-1} exists as an operator on l^1_∞ . The associated lower triangular matrix representation of A, B is

$$A = (a_i(t+j)) \quad B = (b_i(t+j)) \quad j \geq t, \quad t = 0, 1, 2, \dots$$

The operator A^{-1} is well defined if and only if $a_i(0) \neq 0 \forall i$. Equivalently, the process is described as,

$$(A_t y)(t) = (B_t u)(t) \quad \forall t \geq 0$$

where

$$\hat{A}_t(z) = \sum a_i(t) z^i \quad \hat{B}_t(z) = \sum b_i(t) z^i.$$

The linear time-invariant system associated with $\hat{B}_t(z)/\hat{A}_t(z)$ is referred to as the *frozen-time* system.

The above description does not include every possible linear, causal, time-varying system. However, it includes most systems of interest. It follows from an argument similar to those in Francis (1987), that if the time varying plant is stabilizable by a time-varying controller then the plant admits such a representation. In addition, time-varying systems that arise through estimation procedures are automatically described in terms of the above model.

This general description makes it possible to consider infinite-dimensional time-varying systems. Thus, the results of this paper go beyond the standard results that assume finite-dimensional systems, or equivalently, requiring the polynomials $\hat{A}_t(z), \hat{B}_t(z)$ to have *a priori* fixed degree for all t . It is exactly this point that makes the analysis more difficult than the standard state-space analysis.

In many applications, including adaptive control (Dahleh and Dahleh, 1990; Goodwin and Sin, 1984), the operators A_t and B_t are finite pulse response sequences. Such sequences may arise due to an identification procedure to estimate the coefficients in the polynomials that define a rational process model.

2.3. Frozen-time control As was pointed out earlier, the control law is designed on the basis of the frozen-time plants. The design procedure can be effected to perform any number of tasks, for example good tracking and disturbance rejection properties. Each time a controller is designed, the plant is thought of as a linear shift-invariant plant with its defining operators fixed at the values they had at that particular time. To allow ourselves the flexibility of using the controller for several instances in time, we will consider the control design every T time steps. For notational reasons, define $n_t = nT$ where n is the smallest integer such that t lies in the interval $[nT, (n+1)T]$. The controller is designed at the time epochs nT . The methodology by which the controller is designed can vary, and for the sake of keeping the development as general as possible we will not restrict its degree.

Consider the closed loop system depicted in Fig. 1. The closed loop system is said to be stable if the map from u_1, u_2 to y_1, y_2 is bounded (Desoer and Vidyasagar, 1975). The

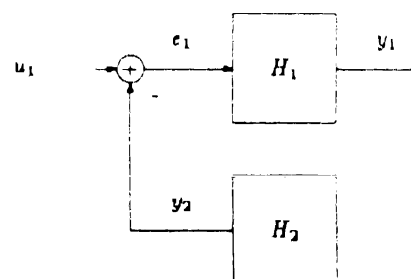


FIG. 1. General form of the closed loop being considered

dynamics of the control law will be given by

$$(L_{n_i} y_2)(t) = (M_{n_i} e_2)(t)$$

where $L_{n_i}, M_{n_i} \in \mathcal{L}_{T1}$ for each index n_i . The evolution of these operators is given by

$$\begin{aligned} (L_{n_i} y_2)(t) &= \sum_{i=0}^t L_{n_i}(t-i) y_2(i) \\ (M_{n_i} e_2)(t) &= \sum_{i=0}^t m_{n_i}(t-i) e_2(i) \end{aligned}$$

The frozen-time operators that define the above control law satisfy the following Bezout identity

$$L_{n_i} A_{n_i} + M_{n_i} B_{n_i} = G_{n_i}$$

where $G_{n_i}^{-1} \in \mathcal{L}_{T1}$ for each fixed n_i . Equivalently, $\hat{G}_{n_i}(z)$, the associated Z-transform, is a polynomial with no zeros inside the closed unit disc. It is important to stress that the degrees of these polynomials may vary for different n_i .

3. Problem formulation and the main result

The main goal of this paper is to show how the slow variation properties of the plant, the controller, the closed loop polynomial interact in order to produce an L^p stable system. The fact that the controller is updated only every T steps introduces a new parameter in the stability analysis. In what follows it is shown how large T can be without endangering the stability of the closed loop system. Intuitively, the larger T is, the slower the plant variation ought to be, and for the extreme case of $T \rightarrow \infty$ the system ought to be time-invariant.

From Fig. 1 we can write down the closed-loop equations for the controlled system as follows

$$(A_i y_1)(t) = (B_i(u_1 + v_1))(t) \quad (1)$$

$$(L_{n_i} y_2)(t) = (M_{n_i}(u_1 + v_1))(t) \quad (2)$$

$$A_{n_i} L_{n_i} + M_{n_i} B_{n_i} = G_{n_i} \quad (3)$$

Next we will obtain a relation that connects the input sequences $\{u_i(t)\}$ and $\{u_2(t)\}$ to the outputs $\{y_1(t)\}$ and $\{y_2(t)\}$. Operating on equation (1) by L_{n_i} we get

$$(L_{n_i} A_i y_1)(t) = (L_{n_i} B_i u_1)(t) = (L_{n_i} B_i v_1)(t)$$

By adding, subtracting, and grouping certain terms we get

$$\begin{aligned} & \{(L_{n_i} A_{n_i} + B_{n_i} M_{n_i}) y_1 + ([L_{n_i}, A_i] + (L_{n_i} A_i - L_{n_i} A_{n_i}) \\ & + [B_i, M_{n_i}] + (B_i M_{n_i} - B_{n_i} M_{n_i})) v_1 \\ & + ([L_{n_i}, B_i] - [B_i, L_{n_i}]) y_2\}(t) \\ & = (L_{n_i} B_i u_1)(t) = (B_i M_{n_i} u_2)(t) \end{aligned}$$

where we used the notation

$$[A_i, B_i] = A_i B_i - A_i B_i$$

To obtain the second closed-loop equation, multiply equation (1) by M_{n_i} :

$$(M_{n_i} A_i y_1)(t) = (M_{n_i} B_i u_1)(t) = (M_{n_i} B_i v_1)(t)$$

Again, by adding, subtracting and grouping we get

$$\begin{aligned} & \{(M_{n_i} B_{n_i} + A_{n_i} L_{n_i}) y_1 + ([M_{n_i}, B_i] \\ & + (M_{n_i} B_i - A_{n_i} B_{n_i}) + [A_i, L_{n_i}] \\ & + (A_i L_{n_i} - A_{n_i} L_{n_i})) v_2 \\ & + ([A_i, M_{n_i}] - [M_{n_i}, A_i]) y_1\}(t) \\ & = (M_{n_i} B_i u_1)(t) = (A_i M_{n_i} u_2)(t) \end{aligned}$$

Define the following quantities for $i = 0, 1, 2$,

$$\begin{aligned} X_i &= [L_{n_i}, A_i] + (L_{n_i} A_i - L_{n_i} A_{n_i}) \\ &+ [B_i, M_{n_i}] + (B_i M_{n_i} - B_{n_i} M_{n_i}), \\ Y_i &= [L_{n_i}, B_i] - [B_i, L_{n_i}], \\ Z_i &= [M_{n_i}, A_i] - [A_i, M_{n_i}]. \end{aligned}$$

$$\begin{aligned} W_i &= [M_{n_i}, B_i] + (M_{n_i} B_i - M_{n_i} B_{n_i}) \\ &+ [A_i, L_{n_i}] + (A_i L_{n_i} - A_{n_i} L_{n_i}) \end{aligned}$$

Using equation (3) we can write the closed-loop equations as follows

$$\begin{bmatrix} G_{n_i} + X_i & Y_i \\ 0 & G_{n_i} + W_i \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} L_{n_i} B_i & R_i M_{n_i} \\ M_{n_i} B_i & A_i M_{n_i} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad (4)$$

Denote by X, Y, Z, W, G the time-varying operators (lower triangular) associated with the families $X_i, Y_i, Z_i, W_i, G_{n_i}$, $i = 0$ respectively. The basic theme of this paper is to analyze the above system by considering the operators X, Y, Z, W as perturbations. These perturbing operators are made up of a sum of simpler commutator type operators. Bounds on the norms of these operators will be derived later.

The main result of this paper is summarized in the theorem below.

Theorem 1. For the system (4) assume the following.

- (AS 1) The operators defining the plant are slowly time-varying with rates γ_A and γ_B , that is $A_i \in \text{SIV}(\gamma_A)$ and $B_i \in \text{SIV}(\gamma_B)$.
- (AS 2) The sequence of controllers are slowly time-varying that is $M_{n_i} \in \text{SIV}(\gamma_M)$, and $L_{n_i} \in \text{SIV}(\gamma_L)$.
- (AS 3) The L_1 norms and the HAH of the operators $A_i, B_i, L_{n_i}, M_{n_i}$ are uniformly bounded in i . This is precisely equivalent to $A, B, L, M, A_i, B_i, L_{n_i}, M_{n_i} \in \mathcal{L}_{1X}$, where A, B, L, M are the time-varying operators associated with the above families and $A_i, B_i, L_{n_i}, M_{n_i}$ are the corresponding derivative operators. From this, AS 1, AS 2, and the Bezout identity, it follows that the closed-loop operator will also be slowly time-varying, and thus we can write $G_{n_i} \in \text{SIV}(\gamma_G)$.
- (AS 4) The L_1 norms and the HAH of the frozen-time linear shift-invariant operators $G_{n_i}^{-1}$ are bounded uniformly in i (this condition will be discussed further in Section 5).

Then there exists a non-zero constant β such that if $\gamma_A, \gamma_B, \gamma_M, \gamma_L, \gamma_G \leq \beta$, the closed-loop system is internally stable.

4. Stability analysis

In this section, the stability of the closed-loop system arising from the frozen-time design is studied. By examining the closed-loop equations (4) we see that the closed-loop transfer function G_{n_i} is perturbed by a few operators, each of which falls into one of the following categories:

- (a) $[A_i, M_{n_i}]$
- (b) $M_{n_i}(A_i - A_{n_i})$

The following lemmas demonstrate how the induced norms of these operators (acting from l^1 to l^1) can be made small by controlling the rates of variation involved in the problem.

Lemma 1. Let $A_i \in \text{SIV}(\gamma_A)$, $M_{n_i} \in \text{SIV}(\gamma_M)$, and R denotes the time-varying operator associated with the commutators $[A_i, M_{n_i}]$, $i = 0, 1, 2$. Then $R \in \mathcal{L}_1$ and its induced norm satisfies

$$\begin{aligned} \|R\|_{\mathcal{L}_1} &= \sup_i \|[A_i, M_{n_i}]\|_{\mathcal{L}_1} \\ &\leq \gamma_M (2T \sup_i \sum_{k=0}^T \|a_i(k)\| + \sup_i \sum_{k=0}^T \|a_i(k)\|) \end{aligned}$$

Proof. We start by showing how the operator $[A_i, M_{n_i}]$ operate on l^1 sequences. Let $v \in l^1$, then

$$[A_i, M_{n_i}]v(t) = \sum_{k=0}^t \sum_{l=0}^k a_i(t-k)(m_{n_i}(k-l) - m_{n_i}(k-l))v(l)$$

Taking the absolute value of the above equation we get:

$$\begin{aligned} \|[A_i, M_{n_i}]v(t)\| &\leq \sum_{k=0}^i |a_i(t-k)| \sum_{i=0}^k \\ &\quad \times \|m_{n_i}(k-i) - m_{n_i}(k-i)\| \|v\|_* \\ &= \sum_{k=0}^i |a_i(t-k)| \sum_{i=0}^k \\ &\quad \times \|m_{n_i}(i) - m_{n_i}(i)\| \|v\|_* \\ &\leq \sum_{k=0}^i |a_i(t-k)| \|M_{n_i} - M_{n_i}\|_{\mathcal{L}_{11}} \|v\|_* \\ &\leq \gamma_M \sum_{k=0}^i |a_i(t-k)| \|n_i - n_k\| \|v\|_* \end{aligned}$$

However,

$$\|n_k - n_i\| \leq \|n_k - k + k - i + i - n_i\| \leq 2T + |k - i|.$$

Therefore, the above inequality can be written as:

$$\|[A_i, M_{n_i}]v\|_* \leq \left(2\gamma_M T \sum_{k=0}^i |a_i(k)| + \gamma_M \sum_{k=0}^i k |a_i(k)| \right) \|v\|_*.$$

Lemma 2 Under the assumptions in Lemma 1, let R denote the time-varying operator associated with the family $M_{n_i}(A_i - A_{n_i}), i = 0, 1, 2, \dots$. Then $R \in \mathcal{L}_{TV}$ and its induced norm satisfies

$$\|R\|_{\mathcal{L}_{TV}} = \sup_i \|M_{n_i}(A_i - A_{n_i})\|_{\mathcal{L}_{11}} \leq \gamma_A T \sup_i \sum_{k=0}^i |m_{n_i}(k)|$$

Proof The proof follows in a similar fashion as above and will be omitted.

Next, the proof of Theorem 1 is presented.

Proof of Theorem 1. Consider the first equation in system (4) expressed in an operator form,

$$Gv_1 + Xv_1 + Yv_2 = v$$

where $v(t) = (L_{n_i} B_i u_1)(t) + (B_i M_{n_i} u_2)(t)$. Letting τ be a fixed time instance ($\tau \in Z^+$), we can write

$$G_{n_i} v_1 + (G - G_{n_i})v_1 + Xv_1 + Yv_2 = v$$

where $G_{n_i} \in \mathcal{L}_{11}$. Denote by H_{n_i} the inverse of G_{n_i} . By assumption (AS-4), $H_{n_i} \in \mathcal{L}_{11}$. Therefore, the above equation can be written as

$$v_1 + H_{n_i}(G - G_{n_i})v_1 + H_{n_i}Xv_1 + H_{n_i}Yv_2 = H_{n_i}v.$$

Evaluating the above operator equation at time τ , the following equation evolving in τ describes the same dynamics of the above equation,

$$\begin{aligned} y_1(\tau) + (H_{n_i}(G - G_{n_i})v_1)(\tau) + (H_{n_i}Xv_1)(\tau) \\ + (H_{n_i}Yv_2)(\tau) = (H_{n_i}v)(\tau) \end{aligned}$$

Similarly we can write

$$\begin{aligned} (H_{n_i}Zv_1)(\tau) + y_2(\tau) + (H_{n_i}(G - G_{n_i})v_2)(\tau) \\ + (H_{n_i}Wv_2)(\tau) = (H_{n_i}w)(\tau) \end{aligned}$$

where $w(t) = (M_{n_i} B_i u_1)(t) + (A_i M_{n_i} u_2)(t)$. Combining the equations above, we get the following closed loop system:

$$(I + F) \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}(\tau) = \begin{pmatrix} H_{n_i} v \\ H_{n_i} w \end{pmatrix}(\tau)$$

where

$$F = \begin{pmatrix} H_{n_i}(G - G_{n_i}) + H_{n_i}X & H_{n_i}Y \\ H_{n_i}Z & H_{n_i}(G - G_{n_i}) + H_{n_i}W \end{pmatrix}.$$

The idea is to show that the time varying operator F mapping ℓ^∞ into itself has an induced norm that is majorized by the rates of variation of the operators A, B, L, M and G . In other words the induced norm of the perturbing operator F can be made less than one by choosing the rates of variation sufficiently small. From the previous lemmas, and the fact that H_{n_i} is uniformly bounded, it is clear that each of the time-varying operators generated from each family of

operators $H_{n_i}X, H_{n_i}Y, H_{n_i}Z, H_{n_i}W$ (indexed in τ), have induced norms that are controlled by the rates of variation $\gamma_A, \gamma_B, \gamma_L, \gamma_M, \gamma_G$. The internal stability will follow from the small gain theorem if we show that the induced norm of the operator $H_{n_i}(G - G_{n_i})$ can be analogously controlled.

Next, a detailed calculation of an upper bound of the norm of the operator $H_{n_i}(G - G_{n_i})$ is given. Let $y \in \ell^\infty$ and the output of the operator is x , then

$$\begin{aligned} x(\tau) &= H_{n_i} \left(\sum_{i=0}^k (g_{n_i}(k-i) - g_{n_i}(k-i))y(i) \right) \\ &= \sum_{k=0}^i \sum_{i=0}^k h_{n_i}(\tau-k)(g_{n_i}(k-i) - g_{n_i}(k-i))y(i). \end{aligned}$$

Taking absolute values we have:

$$|x(\tau)| \leq \sum_{k=0}^i \sum_{i=0}^k |h_{n_i}(\tau-k)| |(g_{n_i}(k-i) - g_{n_i}(k-i))| \|y\|_*.$$

By an argument similar to that given in the proof of Lemma 1 it follows that:

$$\|x\|_* \leq \left(\gamma_G \sup_{\tau \in Z^+} \sum_{k=0}^i |h_{n_i}(\tau-k)| k + 2\gamma_G T \sup_{\tau \in Z^+} \sum_{k=0}^i |h_{n_i}(\tau-k)| \right) \|y\|_*.$$

By assumption AS-4 it follows that there exists constants $C_1, C_2 \geq 0$ such that

$$\|x\|_* \leq (\gamma_G C_1 + 2\gamma_G T C_2) \|y\|_*.$$

With this proof we have shown that the induced norms of all the perturbing operators that comprise F can be made small by choosing the rates of variation sufficiently small. Therefore internal stability follows by an application of the small gain theorem.

5. Discussion of the results

Theorem 1 indicates that if Assumptions 1-4 are satisfied, and if the variations are small enough, the closed loop system will be ℓ^∞ -stable. Assumptions 1-3 are reasonable for the plant, and easy to satisfy for the compensator. Assumption 4, however, is harder to satisfy. Necessarily, this assumption implies that the zeros of $\hat{G}_t(z)$ lie outside a disc of radius $1 + \epsilon$, for some $\epsilon > 0$, the converse, however, is not true. The condition on the zeros can be precisely stated as

$$\inf_{|z| \leq 1} |\hat{G}_t(z)| \geq \delta > 0 \quad \forall t.$$

Hence, from the spectral theory of \mathcal{L}_{11} , G_t^{-1} is in \mathcal{L}_{11} for every fixed t . In addition, the H_∞ -norm of G_t^{-1} is uniformly bounded in t , since:

$$\|G_t^{-1}\|_\infty = \sup_{|z| \leq 1} \frac{1}{|\hat{G}_t(z)|} \leq \frac{1}{\inf_{|z| \leq 1} |\hat{G}_t(z)|} \leq \frac{1}{\delta} \quad \forall t.$$

Note that this does not necessarily imply that $\|G_t^{-1}\|_{\mathcal{L}_{11}}$ is uniformly bounded in t . In the case where the degree of \hat{G}_t is fixed *a priori*, for all t , the uniform boundedness of the $\|\hat{G}_t^{-1}\|_\infty$ will imply the uniform boundedness of $\|G_t^{-1}\|_{\mathcal{L}_{11}}$ (Doyle, 1985).

In the following theorem, we will show that with some mild assumptions on $\hat{G}_t(z)$, the spectral condition is enough to verify the uniform bounds on the \mathcal{L}_{11} -norms and the ITAE of G_t^{-1} .

Theorem 2. Given the following conditions

- 1. $|\hat{G}_t(z)| \leq M_1 \forall |z| \leq 1, \forall t$
- 2. $|\hat{G}_t^*(z)| \leq M_2 \forall |z| \leq 1, \forall t$
- 3. $|\hat{G}_t^*(z)| \leq M_3 \forall |z| \leq 1, \forall t$
- 4. $\inf_{|z| \leq 1} |\hat{G}_t(z)| \geq \delta > 0, \forall t$ (spectral condition)

then the \mathcal{L}_{11} -norm and ITAE of G_t^{-1} are uniformly bounded in t .

Proof. The proof is based on an interesting theorem by Hardy (Hoffman, 1962), which can be stated as follows: Given a function $\hat{R} \in H_\infty$ with

$$\hat{R}(z) = \sum_{k=0}^\infty r_k z^k$$

then, the coefficients r_k satisfy

$$\sum_{k=1}^{\infty} \frac{1}{k} \|r_k\| \leq 2\pi^2 \| \hat{R} \|_1.$$

Hence, to show that $\|G_i^{-1}\|_{\infty}$ is uniformly bounded, we apply Hardy's theorem on $\hat{R}(z) = (d/dz) \hat{G}_i^{-1}(z)$. Note that

$$\frac{d}{dz} \hat{G}_i^{-1}(z) = -\frac{\dot{\hat{G}}_i(z)}{\hat{G}_i^2(z)}.$$

Hence,

$$\| \hat{R} \|_1 \leq \| \dot{\hat{G}}_i \|_1 = M.$$

Let $\hat{G}_i^{-1}(z)$ be given by

$$\hat{G}_i^{-1}(z) = \sum_{k=0}^{\infty} h_k(k) z^k.$$

Then,

$$\hat{R}(z) = \sum_{k=0}^{\infty} k h_k(k) z^{k-1}.$$

Applying Hardy's Theorem on $\hat{R}(z)$,

$$\sum \|h_k(k)\| \leq 2\pi^2 M / \delta.$$

Hence,

$$\|G_i^{-1}\| \leq 2\pi^2 M / \delta^2 + \|h_0(0)\| \leq 2\pi^2 M / \delta + \delta^{-1}.$$

A similar argument works for $\|A_i^{-1}\|$ by considering $\{\hat{G}_i^{-1}\}^T$. The details are omitted.

From Theorems 1, 2 we obtain the following result.

Theorem 3. With Assumptions 1-3, the spectral condition (4), Theorem 2), and the uniform boundedness of the quantities $\|A\|_1, \|B\|_1, \|I\|_1, \|M\|_1$, the closed-loop system (4) is stable.

Proof. The proof is straightforward since the conditions in Theorem 2 will be satisfied from the Bezout identity relating \hat{G}_i to the above quantities.

It is worthwhile noting at this point that standard results in slowly varying systems assume finite dimensionality. In this setup, this is equivalent to restricting A, B, I, M to having a fixed degree. If the coefficients of these functions are bounded in time, and \hat{G}_i has all its zeros outside a disc of radius $1 + \epsilon$, the conditions of Theorem 2 are immediately satisfied, and the closed-loop system is stable. The result in this paper is a generalization of the above to infinite dimensional systems.

6. Conclusions

It is shown that a controller design based on frozen time versions of a slowly time-varying plant stabilizes the plant if the conditions in Theorem 1 are met. The dimensions of both the plant and controller need not be finite, and the controller can be designed at equally spaced instances in time. The infinite dimensionality of the plant or the controller poses technical difficulties in proving stability, and for that purpose, Hardy's Theorem is invoked. The results in this paper will no doubt have an impact on the design of adaptive controllers (Dahleh and Dahleh, 1990; Goodwin and Sin, 1984), as well as the design of gain scheduled controllers (Shamma, 1988).

Acknowledgement. Mohammed Dahleh is supported by Texas A & M University Engineering Excellence Fund, and Munther A. Dahleh is supported in part by the army research office, Center for Intelligent Control, under grant DAAI03-86-K-0131, and in part by NSF, under grant 8810178 ECS.

References

Dahleh, M. A. and J. R. Pearson (1987a). H^2 optimal controllers for MIMO discrete time systems. *IEEE Trans. Aut. Control*, **AC-32**, 314-323.

Dahleh, M. A. and J. R. Pearson (1987b). H^2 optimal compensators for continuous time systems. *IEEE Trans. Aut. Control*, **AC-32**, 889-898.

Dahleh, M. and M. A. Dahleh (1990). Optimal rejection of persistent and bounded disturbances: Continuity properties and adaptation. *IEEE Trans. Aut. Control*, **AC-35**, 687-696.

Desoer, C. A. and M. Vidyasagar (1975). *Feedback Systems: Input-output Properties*. Academic Press, New York.

Doyle, J. C. (1985). Structured uncertainty in control design. IFAC Workshop on Estimation and Control of Uncertain Systems, Boston, MA.

Francis, B. A. (1987). *A Course in H_2 Control Theory*. Springer, Berlin.

Goodwin, G. C. and K. S. Sin (1984). *Adaptive Filtering, Prediction and Control*. Prentice-Hall, Englewood Cliffs, NJ.

Hille, E. and R. S. Phillips (1957). *Functional Analysis and Semi-groups*. American Mathematical Society.

Hoffman, K. (1962). *Banach Spaces of Analytic Functions*. Prentice-Hall, Englewood Cliffs, NJ.

Shamma, J. (1988). Analysis and design of gain scheduled control systems. Ph.D. Thesis EIDS TH 1270, MIT, MA.

Vidyasagar, M. (1985). *Control System Synthesis: A Factorization Approach*. MIT Press, Cambridge, MA.

Willems, J. C. (1971). *The Analysis of Feedback Systems*. MIT Press, Cambridge, MA.

Adaptive Control*

K. J. Åström and B. Wittenmark

Reviewer: D. W. CLARKE

Department of Engineering Science, University of Oxford,
 Parks Road, Oxford OX1 3PJ, U.K.

LIKE THE quest for the philosopher's stone, the dream of an adaptive controller which sustains optimal performance despite changes in the plant and its environment has stimulated many theoretical and algorithmic developments. An early requirement for adaptation was perceived with the new supersonic aircraft, where large dynamic variations are found in passing through different flight regimes, and "Model Reference" methods were the control engineer's response. These were essentially heuristic, involving simple mathematics and many simulations: a typical leading approach of the period was the "MIT rule", in which a controller parameter θ is adjusted according to

$$\frac{d\theta}{dt} = -\mu \nabla_e e^2,$$

where e is the difference between the outputs of the plant and reference model. At first it appeared that these methods could adapt quickly and reliably, but the overconfidence of the 50s with Model Reference Adaptive Systems led to disaster. There was then a retreat to first principles looking for proofs of stability and for new methods with guaranteed behaviour. Applications of MRAS withered, control theory became dominated by the new state-space/LQG methods, and gain-scheduling using pilot pressure to index the appropriate control parameters was used for aircraft control.

Process control lagged behind aerospace developments, with the first Direct Digital Control systems appearing in the 60s, and with DDC universally employing PID algorithms. These sometimes proved difficult to tune, particularly for slow processes or, more significantly, for systems with dead-time. The position was transformed in 1973 by the seminal paper by Åström and Wittenmark which opened up the new field of Self-Tuning Control. Here the ambitions were different: discrete-time rather than continuous-time control [though Gawthrop (1987) develops self-tuning from a continuous-time viewpoint], plants assumed constant (or slowly varying) but unknown rather than rapidly varying dead-time and stochastic disturbances included. This paper, arriving at the time of the energy crisis (and hence the need for improved process efficiencies) and the advent of the microprocessor (giving a cheap route into implementation) aroused intense interest: many new developments and important industrial applications were soon reported. Commercial products became available and there was rapid feedback from practice into new theory (see Harris and Billings, 1981). It was found that in many cases the simple tuning of a fixed-parameter PID regulator was sufficient, rather than continual adjustment of the more complex algorithms provided by STC—a tradeoff between performance and complexity. This sparked off the development of "Auto-tuning" (associated again with the name of Åström).

Meanwhile the MRAS school produced some fundamental stability theory which revolutionised the field. Lyapunov methods were used to "redesign" the MIT rule to give global stability based on the concept of Positive Real transfer functions. Hyperstability and passivity arguments were also

employed. Non-positive real transfer functions were treated using the idea of augmented error. A rapprochement between STC and MRAS was made and there has been significant cross-fertilization between the approaches (Sastri and Bodson, 1989). Though effective designs became available, the essentially nonlinear nature of adaptive controllers means that their behaviour in non-ideal cases is generally difficult to analyse—indeed lack of robustness to unmodelled dynamics has been seen to be a significant problem and fixes such as normalisation, dead-bands, and regressor filtering have been developed. With slow adaptation the two-time-scale nature of the system admits the use of averaging methods (Anderson *et al.*, 1986) so that some general analytical tools have become available.

Adaptive control covers an immense range of subjects with links to many of the ideas of modern control theory. There is no general theory (unlike say for linear systems), but a toolbox of approaches [see, for example, Gupta (1986)] which have proved useful for analysis or application. The choice of which topics to cover is difficult, particularly for an introductory text. The authors deal with this problem with great skill, clearly demonstrating their unrivalled experience in generating and applying many of the leading methods in the field.

Adaptive Control is intended as an introductory course for students who are presumed to have a background in automatic control and sampled-data systems, and is accessible also to an industrial readership. Hence, unlike many books which concentrate just on the theory, the topic is strongly motivated by a discussion of the class of practical problem which might require adaptive control and of the design and use of current commercial adaptive controllers. This discussion includes cases where constant feedback is sufficient and a demonstration that simply inspecting the variations of open-loop step response is *not* sufficient for determining the need for adaptation. Another significant pedagogical feature of the book is the large number of simulations together with SIMNON programs, so that the student is encouraged to repeat the exercises and discover the effect of modifications to the algorithms.

After the introductory motivation, recursive parameter estimation is treated, concentrating on RLS (with forgetting) but with a brief mention of the projection algorithm [certainly much less detailed than in Goodwin and Sin, (1984)] and a discussion of square-root methods for ensuring good numerical properties. The RLS scheme is developed from classical "of-line" least squares followed by application of the matrix inversion lemma. There is a comment about bias when LS is applied to stochastic models and with correlated errors and some brief discussion about ELS/RML, but no demonstration of the magnitude or source of the bias is given, which might worry the non-expert reader. This is the least satisfactory chapter of the book as no simulation is given, nor is Ljung's (1987) useful frequency-response interpretation of least-squares and the role of frequency weighting described. The discussion of the importance of regressor filtering is left to much later in the context of applications, this perhaps is the appropriate place for square-root methods.

In contrast the chapters on MRAS and Self-tuning Regulators are nicely judged, starting with clear descriptions of the basic problems and a methodical development through the increasing complexities of the subjects. For the model reference approach, variants of the MIT rule, Lyapunov and hyperstability methods, positive-real systems,

* *Adaptive Control* by K. J. Åström and B. Wittenmark
 Addison-Wesley, Reading, MA (1989). ISBN 0-201-49720-6

passivity and augmented errors are discussed, leading to a MRAS for a general linear system. Self-tuning is introduced to two simple indirect examples with comments about the difficulty of analysis because, for example, the estimates map into the controller parameters in a complicated way which may have singularities. The "classical" direct methods of minimum-variance and generalized-minimum-variance are developed in an elegant unified style. The more recent indirect approaches of LQG and adaptive predictive control are described.

The book then changes gear by giving an overview of the analysis of stability, convergence and robustness in adaptive schemes. It emphasises that there is only a partial collection of results because of the underlying nonlinear character of adaptive systems. A typical global stability result (using the Key Technical Lemma) is produced for a specific direct pole-placing law for a deterministic minimum-phase plant with known dead-time, and by example shows that certain disturbance patterns added to the plant can produce unbounded behaviour. This motivates the use of dead-zones to recover boundedness. The principal theme of the chapter, however, is the introduction to and use of averaging analysis for slowly-adapting systems. The examples are carefully chosen to give insight into adaptive behaviour: for example it is demonstrated that the SPR rule is significantly less robust than the MFT rule for even simple unmodelled dynamics, with the comment that "analysis of the ideal case can be quite misleading". For stochastic systems there is a heuristic discussion of stochastic averaging leading to Ljung's ODE method, with an illustration of its application to the moving-average self-tuner. There is a brief discussion of instability mechanisms and of Nussbaum's universal stabilizers, though chaotic behaviour in adaptive systems, which has aroused recent interest, is not included. This chapter, together with its successor on the subject of stochastic adaptive control (dual control and suboptimal methods such as cautious control), provides a most interesting summary of the theoretical basis for adaptive control which would prove useful to student and teacher alike.

The next chapters, on auto-tuning and gain scheduling, are novel in a book of this nature and show the applications commitment of the authors. Though relatively simple in theory, the techniques are likely to be of the greatest benefit to industry. The rather brief account of auto-tuning (or pre-tuning) concentrates on the open-loop step response approach, the Ziegler-Nichols method and its more recent development using relay feedback. The basic ideas are clearly explained, but the reader will gain the impression that there are further tricks which need to be up the manufacturer's sleeve in order to design a satisfactory product. Gain scheduling is treated by a series of interesting examples: a nonlinear actuator, tank with variable cross-section, concentration control, a large amplitude pendulum, ship steering, pH control, combustion control, fuel-air control, and flight control. The students will enjoy working through these; the teacher will be reassured that control theory can be useful!

There follows a short chapter on alternatives to adaptive control including robust control, self-oscillating systems, and variable-structure systems. Though this contains interesting material, it is possibly out of sequence and better read in conjunction with the concluding chapter on perspectives of adaptive control. There is little space for issues such as the trade-off between performance and robustness in adaptive control (i.e. distinguishing between robust adaptive control and adaptive robust control ...).

A high-spot of the book is the discussion of practical

aspects of adaptive control, which should be compulsory reading for any practitioner. Signal conditioning (e.g. anti-aliasing filters and their effect on the estimation), the filtering of data against disturbances and unmodelled dynamics, estimator and integral windup, and the interaction between estimation and control are described. A prototype algorithm which combines these features is presented. Many of the recommendations are based on engineering judgment, developed from experience, of the important factors for creating a successful adaptive controller. The experienced reader will be able to read between the lines and recognise the many pitfalls that arise when the advice is ignored; the student may be overwhelmed by the details. The following chapter on commercial systems and applications confirms that extensive built-in knowledge and good user-interface design is essential for worthwhile use of adaptive control.

Adaptive control is a minefield: it is always possible to "break" an adaptive controller by inappropriate use, such as excessive adaptive gain, unmodelled dynamics, non-linear (e.g. hysteresis) actuators, bad disturbance patterns, lack of excitation, etc. This excellent book gives an authoritative guide through these dangers with clear descriptions of the principal techniques and their properties. The breadth of topics covered is very large, so that the student will often need to read supplementary material to obtain detailed knowledge, but the range of examples and illustrations given in the text will surely motivate this effort. The book is destined for every control engineer's shelf.

References

- Anderson, B.D.O., R.R. Bitmead, C.R. Johnson Jr., P.V. Kokotovic, R.L. Kosut, I.M.Y. Mareels, I. Praly and B.D. Riedel (1986) *Stability of Adaptive Systems: Passivity and Averaging Analysis*. MIT Press, Cambridge, MA.
- Astrom, K.J. and B. Wittenmark (1973) On self tuning regulators, *Automatica* **9**, 185-199.
- Gawthrop, P.J. (1987) *Continuous-Time Self-Tuning Control, Vol. 1: Design*. Research Studies Press, Letchworth.
- Goodwin, G.C. and K.W. Sin (1984) *Adaptive Filtering, Prediction and Control*. Prentice-Hall, Englewood Cliffs, NJ.
- Gupta, M.M. (ed.) (1986) *Adaptive Methods for Control System Design*. IEEE Press, New York.
- Harris, C.J. and S.A. Billings (eds) (1981) *Self-Tuning and Adaptive Control: Theory and Applications*. Peter Perigrinus, Stevenage, UK.
- Ljung, L. (1987) *System Identification*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Sastry, S. and M. Bodson (1989) *Adaptive Control*. Prentice-Hall, Englewood Cliffs, New Jersey.

About the reviewer

David Clarke has worked at the Department of Engineering Science in Oxford University for many years, being elected the Reader in Information Engineering in 1986. His research interests lie in the field of self-tuning control and its application using microprocessors, and in the use of signal-processing techniques for local sensor validation. He helped develop the Generalized Least Squares method for identification, the "Clarke-Gawthrop" method in self-tuning, and the recent Generalized Predictive Control approach. He is an Associate Editor of *Automatica* and won an *Automatica* prize in 1987 for his paper on the self-tuning control of non-minimum-phase systems. He was elected to the Fellowship of Engineering in 1989.

Introduction to Signals and Systems*

Edward Kamen

Reviewer: J. F. BÖHME

Ruhr-Universität Bochum, Fakultät für Elektrotechnik,
Lehrstuhl für Signaltheorie, Universitätsstrasse 150, D-4630
Bochum 1, F.R.G.

THE PURPOSE of the book is, following the author in his Preface, to present an introductory yet comprehensive treatment of signals and systems, with a strong emphasis on computing, using programs written in Basic. The book is addressed to students and professionals not only in electrical engineering, but also in mechanical, chemical and industrial engineering, and mathematics. The background needed for reading the book consists of the first and second year courses in calculus, physics and elementary differential equations. A course in electrical engineering is not required, although electrical circuits provide many examples given in the book. Over 250 examples and 280 homework problems help to make the scope of the work quite broad. The derivation of results usually starts with basic facts and is precise, yet mathematical rigor is not overemphasized.

Apart from the broad engineering scope, the author follows the interesting concept to develop continuous-time systems side by side with discrete-time systems. For every model given in the continuous-time case, the corresponding discrete-time system is also studied and *vice versa*, so differences and similarities can be pointed out throughout the book. Additionally, the generation of discrete-time models from continuous-time ones by several methods is discussed in some detail.

Beginning the book with elementary descriptions of signals in the time domain, the author uses input/output differential and difference equations to introduce linear and time-invariant systems. They are illustrated by analog RLC circuits in the continuous-time case and by block diagrams in the discrete-time case. Time-varying, non-linear and multi-input multi-output systems are briefly described. The convolution representation, i.e. the convolution integral and the convolution sum of the system output are introduced. A large effort of the book is devoted to the Laplace and the z transforms, the properties of which are studied and compared. Examples are given as to how to handle linear systems with rational transfer functions. Stability properties of continuous-time systems and discrete-time systems can now be discussed. The asymptotic, marginal, and BIBO stabilities are studied, and conditions and tests of stability are given. The role of frequency response of a system is also studied.

The following topics concern Fourier series as well as Fourier transforms and their properties. Then, system analysis via the Fourier transform is discussed, in particular the output response, amplitude modulation, and pulse amplitude modulation in connection with sampling. The discrete Fourier transform and its properties, system analysis

possibilities and the FFT algorithm are studied. Aspects of analog-to-digital conversions, digital-to-analog conversions, matched filters and digital control are briefly discussed before the book concludes with a chapter on state space representations of linear systems. An Appendix contains the Basic programs and brief reviews of complex variables and matrix algebra. The bibliography lists a selection of related textbooks.

Some remarks on special points of the book are added. First, the "side by side" developments of continuous-time and discrete-time systems are well organized and help the reader to better understand both the concepts and the relationships. I enjoyed reading Chapter 8, in which the three different types of stability (asymptotic, marginal and BIBO) are well described. In particular, the mutual relations are clearly explained, involving little mathematics. The examples are illustrative, as most of them are derived from basic physics and engineering facts; the importance of concepts of material become transparent. However, in my opinion, the examples take up an excessive fraction of the book: the number (over 250) is too many and some could have been developed in less detail. Although the book contains over 600 pages, I missed certain details. For example, solutions of differential or difference equations are only constructed by Laplace transforms or z transforms, respectively. The use of the zeros of characteristic polynomials is not indicated. The author uses deterministic signals and systems throughout the book. A short discussion of stochastic signals to describe noise phenomena would be helpful. The author indicated a strong emphasis on computing using programs written in Basic in his Preface: the 14 Basic programs included, however, are weakly documented and cover only a limited set of problems. Also, using Basic to write scientific and engineering programs is out of date. Finally, I think that most readers do not require authors' programs, because there is a large number of professional software packages available that run on PCs and that do the same tasks and much more.

In the reviewer's opinion, the book is clearly written and easy to understand. The author has, in general, gained his ends. This textbook can be recommended to engineering students in the second or third year. For instructors, there is also a well selected schedule of coverage for a one semester course of 40 lectures of 50 minutes duration. The book will find its place within the big family of different books on signals and systems and signal processing.

About the reviewer

Professor Böhme obtained the Diploma in mathematics from the Technical University of Hannover in 1966 and the Dr.-Ing. in 1970 from the University of Erlangen and the Habilitation in 1977 from the University of Bonn, both in computer science. He was with Krupp-Atlas Elektronik in Bremen and with EGAN in Wachtberg-Werthhoven. Since 1980, he has been Professor of Electrical Engineering at Ruhr University Bochum. His research interests are in array signal processing, detection and estimation, parallel algorithms and systolic arrays. He is a Fellow of the IEEE.

* *Introduction to Signals and Systems* by E. Kamen
Macmillan, New York (1987) ISBN 0-02-362950-9,
U.S. \$36.00.

State Variable Methods in Automatic Control*

K. Furuta and A. Sano

Reviewer: J. L. WILLEMS

Rijksuniversiteit Gent, Groote Steenweg Noord 2, B-9710 Gent, Belgium.

It is HARD to characterize the objectives of this book compared to the many other books available on linear systems (e.g. Kwakernaak and Sivan, 1972; Kailath, 1980; Knobloch and Kwakernaak, 1985). The authors are not helpful in this respect since the book does not contain a preface.

The book does not intend to offer new viewpoints or original results. It contains more or less the material which is generally taught at the senior undergraduate level or first year graduate level in a course on linear control systems theory. Although the title does not say so, the book only deals with linear continuous-time time-invariant (control) systems. The following topics are treated: impulse response, transfer function, state models, differential equations and the relationships between them; controllability, observability, and the corresponding system decomposition; (minimal) realizations and algorithms to obtain them; state feedback, pole assignment and decoupling; observers; LQ-optimal control, Kalman filtering and LQG-stochastic optimal control.

I have the feeling that this book is not well suited for independent reading. There is very little motivation for the various topics discussed. Moreover the reader is not told where the analysis is rigorous and where it is not. In addition sometimes the analysis is not rigorous, even where the reader might think it is.

A few examples.

— The controllability property is defined with respect to the possibility of transferring any initial state at time zero to the zero state at time t_1 . No comments are given on the obvious questions: is t_1 given, is it arbitrary or should the property hold any t_1 ? Why only steer to the zero state? The former remark is relevant if one thinks of time-varying systems, the latter if one looks at discrete-time systems. The condition for controllability is expressed in terms of the nonsingularity of the Gramian for t_1 ; here the same questions obviously arise. Similar remarks can be made with respect to the definition and the analysis of observability.

— In the analysis of the LQG-problem the concepts of optimality and linear-optimality are mixed up.

* *State Variable Methods in Automatic Control* by K. Furuta and Akira Sano. Wiley, Chichester (1988). ISBN 471918776, £16.95.

— For the LQ-optimal control problem, it is not obvious that the limiting case of the receding horizon optimal control problem corresponds to the solution of the infinite horizon optimal control problem. This has been discussed in depth by Callier and Willems (1981) and Willems and Callier (1983).

I have serious objections to the way Lyapunov theory is applied. The authors do not seem to see the distinction between negative definiteness of the derivative of the Lyapunov function on the one hand and the property that the derivative is negative semi-definite but only identically zero along the null solution, on the other hand. If the matrix H is not square, then $-||Hx||^2$ cannot be negative definite (page 75).

Another weak point of the book is the list of references. The most recent reference is more than 10 years old, this is very surprising in a field where much new insight and many new results have been obtained during the last decade.

My conclusion is that this book is not really suitable for independent study. It may be useful for a lecturer teaching a course on linear control systems theory. For him, the main interesting features of the book will be: the selection of the topics discussed, the numerical examples and the exercises. The mathematics involved are much simpler than in many recent texts on linear systems with heavy emphasis on geometric control theory. This may be an asset if one wants to use the book as complementary reading material for a course in an engineering school.

References

- Callier, F. M. and J. L. Willems (1981). Criterion for the convergence of the solution of the Riccati differential equation. *IEEE Trans. Aut. Control*, **AC-26**, 1232–1242.
- Kailath, T. (1980). *Linear Systems*. Prentice-Hall, Englewood Cliffs, N.J.
- Knobloch, H. W. and H. Kwakernaak (1985). *Lineare Kontrolltheorie*. Springer, Berlin.
- Kwakernaak, H. and R. Sivan (1972). *Linear Optimal Control Systems*. Wiley-Interscience, New York.
- Willems, J. L. and F. M. Callier (1983). Large finite horizon and infinite horizon LQ-optimal control problems. *Opt. Control Applic. Meth.*, **4**, 31–45.

About the reviewer

J. L. Willems is with the University of Gent, Belgium, where he is a Professor of Electrical Engineering and Dean of the Engineering Faculty. His main research and teaching interests include linear systems theory, decentralized and stochastic control systems, and power system analysis.

Binäre Steuerungstechnik—Eine Einführung*

K. H. Fasol

Reviewer: R. JOHANSSON

Lund Institute of Technology, School of Technical Physics, Department of Automatic Control, Lund, Sweden.

AUTOMATION TECHNOLOGY may cover at least three meanings of "control" which apart from feedback control may denote the important areas of sequential control and

supervisory control. Manufacturers of industrial control systems have lately incorporated many of these aspects into their products, and thus there appear new demands that application engineers should master several new aspects of control. The countries of the European community seem to have acquired a high state of maturity with achievements such as Grafcet, and industrial standards for sequential control. Adequate engineering education at a lower and intermediate level, and thus good textbooks, are obviously prerequisites to the broad application of these new means of technology.

* *Binäre Steuerungstechnik—Eine Einführung* by Karl Heinz Fasol. Springer, Berlin (1988).

One book in this category is *Binäre Steuerungstechnik—Eine Einführung*. It tries to fill a gap between the theoretical approaches of computer science and the practical attitudes of automation technology with a special attention to the needs of mechanical engineers. The book consists of two parts covering combinatorial circuits (Chapters 1–6) and sequential circuits (Chapters 7–10), the latter in turn divided into sequential machines and programmable controllers. The introduction states the advantages and disadvantages of programmable and non-programmable solutions, respectively.

The first part of the book describes combinatorial design with Boolean algebra, predicate calculus and set theory. Graphical representations such as Karnaugh diagrams, contact charts, function block analysis and ladder diagram languages are presented in a traditional way. Minimization of logic circuits *ad modum* Quine, McCluskey, Karnaugh and the conjunctive minimal form are treated in detail but illustrations in the form of Venn diagrams are absent. Circuit design with implementation by NAND- and NOR-elements and the problems of hazard are presented in a traditional way.

Sequential machines are thoroughly presented with detailed functional block diagrams including gates and flip-flops. Some design examples are presented with combinatorial and sequential circuits connected to the clock pulse input, although such a method is sometimes considered as poor design practice. The attention to logic design problems here dominates the problems of interest for programmable logic controllers. There is a comparison (p. 101) between hardwired (Verbindungsprogrammierung) and programmed (Speicherprogrammierung) and other sequential machines but the section contains no traditional classification such as Mealy and Moore machines. Special cases of finite state machines in the form of sequential function charts are covered in detail.

The distinction between single sequential function charts (Zwangstolgesteuerung) and methods allowing sequence selections (Freifolgesteuerung) is carefully pointed out in Chapter 8.

Chapter 9 treats Huffman automata, with special attention to the hazards and oscillations appearing in asynchronous circuits, and the special coding methods necessary for hazard-free design of asynchronous sequential machines. Limitations of the Huffman finite state machines as to the simultaneous synchronization and sequencing are also reviewed. Remedies by means of coding are presented but

the practically important Gray codes have been omitted.

The presentation style and the focus of attention change considerably between Chapters 9 and 10. Chapter 10 introduces the programmable controllers with historical remarks and with comments on the importance of microcomputer implementations. Several similarities to computer programming are emphasized with recommendations as to programming style, modularization and programming environments.

The application examples are centered around a Siemens programmable control system. The book also contains a valuable appendix that includes several exercises that make it suitable for course work.

Historical remarks are, with few exceptions, given only in the introduction which emphasizes the change of focus from pneumatic and electromechanical devices to electronic devices.

References to industrial standards such as DIN, VDI and IEC have been made in a careful way. However, references to original work are not quite systematic in the text. In particular, contemporary work including the author's own work has been given much attention in the reference list.

There is obviously a considerable difference between the needs of logic design (e.g. VLSI design) and the asynchronous logic circuits necessary for automation and application problems of mechanical engineering. This accounts for some incoherence in the presentation style which is most obvious to the reader in Chapters 9 and 10.

As an overall evaluation, it may be concluded that this book is carefully presented and illustrated with many examples. Bearing in mind the strong differences in attitudes between the schools of technology and the universities in the educational systems of central Europe, this textbook probably satisfies the need for a comprehensible textbook for a broader and practically oriented engineering student community. It may also be recommended as an introductory textbook of automation for intermediate level engineering curricula. Among the deficiencies, it may be mentioned that there is no English vocabulary and there is only one standard set of logic design symbols. This is obviously in the spirit of the close references to industrial standards but may certainly present a problem to a student of scientific and technical literature in English and may put a limitation on the usefulness of the book. Secondly, there are no illustrations in the form of state transitions graphs, which may be a disadvantage for pedagogical purposes.

Statistical Analysis and Control of Dynamic Systems*

H. Akaike and T. Nakagawa

Reviewer: RUDOLF KULHAVÝ

Czechoslovak Academy of Sciences, Institute of Information Theory and Automation, Pod vodárenskou věží 4, 182 08 Prague 8, Czechoslovakia.

THIS RELATIVELY short book does not belong to a group of traditionally treated textbooks on statistical analysis or stochastic control although its title could indicate this. Its goal is to communicate the experience of the authors with the application of general theory to a specific example of a complex technological process—a cement rotary kiln. There are a great many publications on specified topics of theory today, but a book trying to enlighten the nature of problems

encountered during the implementation phase, through a very detailed case study, is still rather exceptional.

This was very probably the reason to introduce this book in the Japanese series of the programme "Mathematics and Its Applications". We must agree with the observation of Prof. M. Hazewinkel in the Series Editor's Preface that "the kind and level of sophistication of mathematics applied in various sciences has changed drastically in recent years". The same is true also for the area of noisy dynamical systems and their control, where both powerful general theory and effective computer algorithms are now available. It could seem that it remains just to apply the ready tools. However, the authors of the book demonstrate in a convincing manner that this last step is far from being straightforward or trivial and deserves much more care than could be expected.

It is of real interest to localize the main source of difficulties. The authors notice that in contrast to e.g. Newtonian mechanics where abstract concepts can be

* *Statistical Analysis and Control of Dynamic Systems* by H. Akaike and T. Nakagawa. Kluwer Academic Publishers, Dordrecht (1989). ISBN 9027727864. \$99.00, £54.00.

identified with their real counterparts uniquely, statistically-based theories are much more sensitive to misinterpretation of their basic concepts. The theory provides only a framework for coherent reasoning, i.e. when applied to an object satisfying its premises, it produces a valid conclusion. It means that before using such a theory, the user must first establish the correspondence between the theoretical model and the real process. Underestimating this point may form a barrier between theory and practice. What can help to avoid the barrier is proper statistical analysis of observational data in combination with a deep understanding of the physical or chemical nature of the process studied. Obviously, this stage is unthinkable without a team cooperation of people having specific knowledge of the process and people skilled in statistical processing of real data.

The authors continuously support their arguments by experience from design of computer control of cement kilns. The cement kiln serves to them only as an example of a system that exhibits a complex behaviour of mutually dependent variables; similar features are typical for many other systems.

Five chapters of the book deal with all the main prerequisites for a successful application. Chapter 1 "What Is the Problem?" explains the need for a thorough prior statistical analysis of any dynamic system, the behaviour of which is to be predicted and/or controlled. Specifically, structure identification appears to be a very important point, as with data of a finite length we must constrain the range of possible structures of the system model. While in ordinary textbooks the statistical structure of the system and the noise sources are assumed known and fixed, in actual problems the decision lies with the user. A detailed analysis may sometimes lead to a modification of the system structure (if this is not rigidly specified). This confirms the central role of a human observer at this stage.

Chapter 2 "An Explanation of the Controller Design Problems" characterizes difficulties with the implementation of control of a stochastic dynamic system. The cement kiln process is described here in detail including the principle of its operation and basic chemical reactions. Process variables and their measuring points are also discussed. However, the main purpose of the chapter is to justify the choice of the method discussed further. To overcome existing difficulties, the authors recommend the following procedure. First, it is necessary to understand the kiln as a statistical dynamic system. The need for respecting statistical fluctuations of the system was totally underestimated in the conventional controller design. Second, the statistical characteristics of the system have to be confirmed through a careful statistical analysis. Third, to design a controller for a multivariate system, the dynamic programming approach should be applied using the state space representation of the system. This can reduce the time and memory requirements of the final algorithm substantially.

Chapter 3 "Statistical Preliminaries" provides in a concise form a self-contained introduction to linear time series analysis and control design. Spectral analysis of stationary time series, together with the task of statistical estimation of spectra is treated. The problem of fitting an autoregressive model to real data is solved including the determination of the model order using the final prediction error method. Then the optimum controller design under a quadratic criterion is described. The authors provide *ad hoc* rules for the choice of the penalization matrices entering the performance criterion. A possible extension of theory to the mixed autoregressive moving average model is discussed. Its use could decrease the number of coefficients to be estimated but the difficulty of numerical calculation makes this approach still less favourable.

Chapter 4 "A Successful Application" reports the results achieved with the help of the described methodology. Technical details are avoided [the interested reader may find them in Otomo *et al.* (1972)] to stress generally valid conclusions. The reader may inspect some results of prior spectral analysis, e.g. compared power spectra of controlled signals, or coherencies between the kiln end gas temperature and manipulated variables. Special care is taken to the

problem of selection of variables to be manipulated and especially of those to be controlled. Then experience gathered with the final realization of the control system is summed up. In concluding remarks the authors emphasize that the basis of their success was the establishment of a close cooperation between the groups of people concerned with management, instrumentation and control, and statistical methodology. No doubt, this observation has a much more general validity.

The last Chapter 5 "Computer Programs" lists the programs in the FORTRAN language that give the user the possibility e.g. to compute the estimate of the auto-covariance or cross-covariance function from a given set of data, to evaluate the estimate of a spectral density function, to fit an autoregressive model to a real process, to compute spectra through autoregression, or to perform optimal controller design. A collection of these programs took the form of a program package for time series analysis and control (TIMSAC). To provide a brief review of important applications as well as a further development of the TIMSAC package, a survey paper originally published in the Bulletin of the International Statistical Institute is reproduced in the Appendix.

The original Japanese version of the book was published in 1972. Since that time significant progress has been made both in theory and software. Let us mention at least the formulation of the AIC (Akaike's information criterion) procedure for the determination of the order of an autoregressive model that made the first of the authors extremely well known in the field. Nevertheless, the methodological communication of the book remains alive.

Note that in comparison with the early 1970s when the original version was written there is more optimism today as concerns partial automation of the implementation phase. In fact there is now a good algorithmic support e.g. for the determination of model orders and delays, or for the selection of relevant variables. Other problems such as the choice of the sampling period, or the adjusting of the penalization matrices in the quadratic criterion are under study. A tremendous progress in software technology during the last 15–20 years is useless to mention. The use of computer aided design tools can dramatically reduce the implementation costs. In spite of these facts, to eliminate fully the creative role of highly qualified specialists at this stage will be scarcely ever possible.

The book brings a lot of convincing facts that a successful application is a result of combination of more factors, mainly the availability of powerful theoretic tools and effective computer programs, a deep understanding of the physical nature of the process under study, and rational management and organization of the whole effort. Thus, the book is attractive both for practising engineers who may get a better view of contributions that modern statistical and control theories can bring to them, and for researchers who may become more aware of the nontrivial nature of the actual implementation of theoretical ideas.

References

- Otomo, T., T. Nakagawa, and H. Akaike (1972). Statistical approach to computer control of cement rotary kilns. *Automatica*, 8, 35–48.

About the reviewer

Rudolf Kulhavý was born in České Budějovice, Czechoslovakia on 4 April 1957. He obtained his first degree (corresponding to the M.Sc. degree) from the Czech Technical University, Faculty of Electrical Engineering in 1981, and the degree of Candidatus Scientiarum (corresponding to the Ph.D. degree) from the Czechoslovak Academy of Sciences, Institute of Information Theory and Automation in 1985. Since 1985 he has been on the research staff of the Department of Automation headed by Dr V. Peterka. His current research interests include real-time identification of non-stationary, non-linear, or non-Gaussian systems, self-tuning control, and computer-aided design of control systems.

Engineering Applications of Stochastic Processes: Theory, Problems and Solutions*

Alexander Zayezdny, Daniel Tabak and Dov Wulich

Reviewer: J. MICHÁLEK

Czechoslovak Academy of Sciences, Institute of Information Theory and Automation, Pod vodárenskou věží 4, CS-182 08 Prague 8, Czechoslovakia.

THIS BOOK was published in the series *Applied and Engineering Mathematics* and is dedicated to specialists in engineering-oriented disciplines. As is said in the foreword, the main purpose is to give a systematic presentation of the theoretical and practical basic notions of probabilistic calculus and to show close connections with disciplines, such as electronic communication, radar and automatic control. The book is intended for undergraduate and graduate students and practical engineers specializing in the areas mentioned above. It might be used as a self-study text or a text-book for courses on applications of random processes in technical areas.

The authors chose a very difficult task to write a book suitable both for practice and for students, in a field as large as utilization of probability theory and stochastic processes in engineering applications. With respect to this fact it is necessary to mention the contents of the book to show which parts of the theory of probability and random processes were chosen by the authors as most important from the view of practice.

The book is subdivided into two parts: *Random Variables* and *Random Processes*. The first part presents chapters dedicated to the probability theory, the other touches the most important cases of random processes which can be met in practice.

Chapter 1 introduces the basic concepts and definitions of probability, the basic rules of calculating with probabilities and the conditional probability and the Bayes formula. The notion of mutually independent trials and the de Moivre-Laplace theorem are explained in Chapter 2. Chapter 3 deals with the notion of a random variable and its probability distribution function. The characteristic function is also introduced. The notion of entropy, very important in statistical communication methods, is discussed too. Chapter 4 gives a solution of a very important problem, related to nonlinear networks, finding the probability distribution of the function of a random variable. This topic is continued in Chapter 11. Chapter 5 is devoted to the multidimensional vector—case of random variables with a special emphasis on the two-dimensional case. Chapter 6 deals with probability theory, and introduces the laws of large numbers.

Chapters 7–11 form the second part of the book, which is devoted to random processes. The notion of a random process, correlation and spectral analysis are presented in Chapter 7. Chapters 8 and 9 are devoted to the canonical representation of random variables and random processes. Chapter 10 deals with linear systems and their responses to random inputs. The classical case with a time-invariant system and a weakly stationary input is considered. Chapter 11 is the largest chapter and discusses some nonlinear systems and their behaviour under random inputs. Two examples are presented: the inertia-less system and the inertial one—the main topic of this chapter. A variety of methods for overcoming troubles with the nonlinear character of such systems are discussed in detail. An Appendix offers special tables of probability distributions and their numerical characteristics.

After getting acquainted with the contents of the book one could, of course, dispute with the authors their choice of cases useful for practice. However, we must realize the practical impossibility of considering all the examples of random processes related to engineering practice. The authors were also limited by the mathematical means which are used.

The book can be read and studied with only a basic knowledge of higher mathematics, so can mainly be recommended to undergraduate students at technical universities. For graduate students and for engineers who are more thoroughly acquainted with some mathematical disciplines, such as measure theory and abstract integration, the reviewer would recommend books based on a deeper mathematical background. A modern approach to probability theory without the Lebesgue integral is almost impossible. Moreover, the general theory of the Lebesgue integral is simple and more suitable because one need not distinguish between continuous and discrete random variables as is done in the present book. At the beginning, by the presentation of basic properties of probability, the authors could mention in more detail the axiomatics of probability theory and start with a basic triple (Ω, σ, P) based on the notion of σ -algebra of random events, some denotations of operations with random events are nonstandard and somewhat unhelpful. In the opinion of the reviewer, more attention could be paid to the weak stationarity, the most important case of random process related to the practice, e.g. the Wiener-Chinvin theorem should be introduced in an explicit form. With regard to used mathematical tools, some notions and problems are only outlined, and many times accompanied by numerous examples. A typical example is Chapter 8 concerning the canonical expression of the second order random process, which can be described and explained in detail by means of the stochastic integral. Chapters 10 and 11 are presented in a similar way. From this fact, it should be convenient to bring a richer list of suitable references which could be at disposal for more thorough study.

After the critical words, let's emphasize affirmative sides of the book. The reviewer sees its main advantage in a large number of solved examples at the end of every chapter, which occupy approximately one third of the book. In this way the book can also be attractive for specialists in probability theory and stochastic processes.

We can summarize the book as useful for students and engineers in practice, for the first acquaintance with the theory of random processes and with elements of the probability theory. Reading the book carefully, together with the worked examples, is also recommended. This advice follows from the reviewer's conviction that the given examples illustrate utilization of theoretical results in practical situations. The examples can also be suitable for lectures in applied fields of probability and stochastic processes.

About the reviewer

Jiří Michálek was born in Czechoslovakia in June 1943. He graduated in 1966 from the Charles University of Prague, Faculty of Mathematics and Physics. In 1967 he joined the Department of Information Theory at the Institute of Information Theory and Automation of the Czechoslovak Academy of Sciences. He received in 1975 the CSc (Ph.D.) degree in probability theory and mathematical statistics. His research concerns problems of estimation and recognition in random processes, spectral decomposition of nonstationary processes, detection of changes in behaviour of random processes and applications in practical problems.

* *Engineering Applications of Stochastic Processes: Theory, Problems and Solutions* by A. Zayezdny, D. Tabak and D. Wulich. Wiley, Chichester. ISBN 0-863800769, £54.50.

Biographical Notes on Contributors to this Issue



Joergen Ackermann was born in Bochum, Germany, in 1936. He received the M.S. degree in electrical engineering from the University of California, Berkeley, in 1964, and the Dr.-Ing. degree from the Technical University in Darmstadt, in 1967. Since 1974 he has been director of the Institute for Flight Systems Dynamics at the German Aerospace Research Establishment

(DLR) in Oberpfaffenhofen. He is also an adjunct professor at the Technical University in Munich. He has held visiting positions at the Universities of Illinois (Urbana-Champaign) and California (Irvine). He is author of the book *Sampled-data Control Systems* (Springer, 1985) and associate editor at large of the *IEEE Transactions on Automatic Control*. His research interests are in robustness analysis and design of control systems.



James Lenton Alty is the Executive Director of The Turing Institute and, since 1982 Professor of Computer Science at the University of Strathclyde. He received his BSc, with first class honours, and a PhD in nuclear physics from the University of Liverpool in 1961 and 1966, respectively. From 1968 to 1972 he was employed by IBM (U.K.) as a Senior Systems Engineer and

Account Executive, consecutively. He then returned to the University of Liverpool in 1972 as Director of the University Computer Centre. Professor Alty has specialised in Man-Machine Interaction since 1976 and published (with M. J. Coombs) *Computing Skills and the User Interface* (Academic Press, 1981) and *Expert Systems: Concepts and Examples* (NCC, 1984). His research interests include human-computer interaction and knowledge-based systems. He has given seminars and courses in artificial intelligence and human-computer interaction in Europe, U.S.A. and Australia, and has undertaken extensive consultancy with industry.



Michael Athans was born in Drama, Greece on 3 May 1937. He received his B.S.E.E. in 1958 (with highest honors), his M.S.E.E. in 1959, and Ph.D. in control in 1961, all from the University of California at Berkeley.

From 1961 to 1964 he worked at the MIT Lincoln Laboratory, Lexington, MA. Since 1964 he has been with the MIT Electrical

Engineering and Computer Science Department where he is Professor of Systems Science and Engineering. He also

served as director of the MIT Laboratory for Information and Decision Systems (formerly the Electronic Systems Laboratory) from 1974 to 1981. He is co-founder of ALPHATECH, Burlington, MA, where he is Chief Scientific Consultant and Chairman of the Board of Directors, and has also consulted for other industrial organizations and government panels.

Dr. Athans is co-author of *Optimal Control* (McGraw-Hill, 1966), *Systems, Networks, and Computation: Basic Concepts* (McGraw-Hill, 1972), and *Systems, Networks, and Computation: Multivariable Methods* (McGraw-Hill, 1974). In 1974 he developed 65 TV lectures and accompanying study guides on *Modern Control Theory*. He has authored or co-authored over 250 technical papers and reports. His research interests span the areas of system, control, and estimation theory and its applications to the fields of defence, aerospace, transportation, power, manufacturing, economic and command, control, and communications (C³) systems.

He has received many awards including the American Automatic Control Council's 1964 Donald P. Eckman award for outstanding contributions to the field of automatic control, the 1969 F. E. Terman award of the American Society for Engineering Education as the outstanding young electrical engineering educator, and the 1980 Education Award of the AACC for his outstanding contributions and distinguished leadership in automatic control education. In 1973 he was elected Fellow of the IEEE and in 1977 Fellow of the American Association for the Advancement of Science. He has served on numerous committees of IFAC, AACC and IEEE; he was president of the IEEE Control Systems Society from 1972 to 1974. In addition he is a member of Phi Beta Kappa, Eta Kappa Nu and Sigma Xi. He was associate editor of the *IEEE Transactions on Automatic Control*, co-editor of the *Journal of Dynamic Economics and Control* and associate editor of *Automatica*.



Dimitri P. Bertsekas was born in Athens, Greece, in 1942. He received a combined B.S.E.E. and B.S.M.E. from the National Technical University of Athens, in 1965, the M.S.E.E. degree from George Washington University in 1969 and the Ph.D. degree in system science from the Massachusetts Institute of Technology in 1971.

Professor Bertsekas has held faculty positions with the Engineering Economic Systems Dept. Stanford University (1971-1974) and the Electrical Engineering Dept. of the University of Illinois, Urbana (1974-1979). He is currently Professor of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. He consults regularly with private industry and has held editorial positions in several journals. He was elected Fellow of the IEEE in 1983.

Professor Bertsekas has done research in the areas of estimation and control of stochastic systems, linear, nonlinear and dynamic programming, data communication networks, and parallel and distributed computation, and has written numerous papers in each of these areas.

He is the author of *Dynamic Programming and Stochastic Control*, (Academic Press, New York, 1976); *Constrained*

Optimization and Lagrange Multiplier Methods, (Academic Press, 1982); *Dynamic Programming: Deterministic and Stochastic Models* (Prentice-Hall, Englewood Cliffs, NJ, 1987); and co-author of *Stochastic Optimal Control: The Discrete-Time case* (Academic Press, 1978); *Data Networks* (Prentice-Hall, 1987), and *Parallel and Distributed Computation: Numerical Methods* (Prentice-Hall, 1989).



Christopher I. Byrnes was born in New York, on 28 June 1949. He received the B.S. degree from Manhattan College, Bronx, NY, in 1971 and the M.S. and Ph.D. degrees from the University of Massachusetts, Amherst, in 1973 and 1975, respectively.

He served as an Instructor in the Department of Mathematics at the University of Utah, Salt Lake City, from 1975 to 1978,

when he was appointed Assistant Professor in the Department of Mathematics and in the Division of Applied Sciences at Harvard University, Cambridge, MA. From 1982 to 1985 he was Associate Professor of Applied Mathematics on the Gordon McKay Endowment at Harvard University. In 1984, he joined Arizona State University, Tempe, as a Research Professor of Engineering and Mathematics. He is currently Chairman and Professor in the Department of Systems Science and Mathematics at Washington University, St. Louis, MO, and Adjunct Professor of Mathematical System Theory at the Royal Institute of Technology (KTH), Stockholm, Sweden. Dr. Byrnes has also held visiting positions at Bremen, Groningen, Harvard, IIASA, Kansas, KTH, Osaka, Paris Dauphine, Rome-La Sapienza, Stanford and Tokyo Universities. Editor of eleven research volumes and author of over 100 technical articles, his research interests include adaptive control, algebraic system theory, distributed parameter systems, linear multivariable control, nonlinear control, and the applications of nonlinear dynamics in control and estimation. A member of AAAS, AMS, IEEE and SIAM, Dr. Byrnes was named a Case Centennial Scholar by Case Western Reserve University in 1980, a Fellow of the Japan Society for the Promotion of Science in 1986 and the Graduate School Distinguished Research Professor at Arizona State University in 1988. In 1989 he was elected Fellow of the IEEE.

Dr. Byrnes has served as an Associate Editor of six journals and is currently Editor of the two new book series, *Systems and Control: Foundations and Applications* and *Progress in Systems and Control* published by Birkhauser, Boston, MA.



Stephen L. Campbell received the B.A. degree in mathematics from Dartmouth College, Hanover, New Hampshire, U.S.A. in 1967, and the M.S. and Ph.D. degrees in mathematics from Northwestern University, Evanston, IL, U.S.A. in 1968 and 1972, respectively. In 1972 he joined the Department of Mathematics, North Carolina State University, Raleigh, NC, U.S.A. as an

Assistant Professor, becoming an Associate Professor in 1976, and a Professor in 1981. His current research focuses on the numerical and analytical solution of implicit systems

of ordinary differential equations and their applications to control, circuit theory, and mechanics.

Dr. Campbell is a member of the Society for Industrial and Applied Mathematics and its Control and Linear Algebra Activity Groups.



Ye-Hwa Chen was born in Taiwan. He received his B.S. degree in chemical engineering from the National Taiwan University in 1979, and the M.S. and Ph.D. degrees in mechanical engineering from the University of California, Berkeley, in 1983 and 1985.

From 1986 to 1988 he served as a faculty member in the Department of Mechanical and Aerospace Engineering of Syracuse University. He then became a faculty member of the School of Mechanical Engineering of Georgia Institute of Technology. His research has been on advanced control methods for manufacturing systems, adaptive robust control of uncertain systems, large-scale systems and discrete events. Dr. Chen has established and served as the chairman for the Complex Systems Technical Panel of ASME since 1987. He is a member of Sigma Xi.



Emmanuel G. Collins, Jr. was born in Monrovia, Liberia, in 1959. He received the Interdisciplinary B.S. degree from Morehouse College, Atlanta, GA, in 1981, the B.M.E. degree in mechanical engineering from the Georgia Institute of Technology, Atlanta, in 1981, the M.S. degree in mechanical engineering from Purdue University, West Lafayette, IN, in 1982, and the Ph.D.

degree from the School of Aeronautics, Purdue University, West Lafayette, IN, in 1987.

In June 1987 he joined the Structural Control Group at the Government Aerospace Systems Division, Harris Corporation, Melbourne, FL. His current research interests include robust analysis and control and the use of homotopic continuation methods for reduced-order control law synthesis with an emphasis on application to large flexible space structures.



Mohammed Dahleh was born in Jordan in 1961. He received the B.S. degree in electrical engineering from Texas A & M University in 1983, and the M.A. and Ph.D. degrees from Princeton University, in 1986 and 1987 respectively.

Since 1987 he has been with Department of Electrical Engineering at Texas A & M University. His general area of interest is system and control theory; his current areas of interest include robust control, adaptive control and distributed parameter systems.



Matthew A. Dahleh was born in 1962. He received the B.S. degree in electrical engineering from Texas A & M University in 1983, and his Ph.D. degree in electrical engineering from Rice University, in 1987. He is currently an Assistant Professor of Electrical Engineering at the Massachusetts Institute of Technology. He has held consulting positions with NASA and C.S.

Draper Laboratory since 1988. His current interests include robust control, identification of uncertain systems and adaptation, control of time-varying and infinite-dimensional systems.

Dr Dahleh received the George Axelby Outstanding Paper Award with J. B. Pearson in 1989.

Hüseyin Demircioğlu was born in Ankara, Turkey in 1961. He received his B.Sc. degree in electronic engineering from Hacettepe University, Ankara in 1983. From 1983 to 1985, he worked as a research and design engineer for ASELSAN (Military Electronic Industry), Ankara. He has recently obtained his Ph.D. degree from Glasgow University, U.K. His current research inter-

ests include adaptive and self-tuning control, system identification, optimal and predictive control.

Jean-Michel Dion was born in La Tronche, France, in 1980. He graduated in mathematics in 1972. He received the "Thèse de 3ème cycle" and "Thèse d'Etat" degrees both from the Institut National Polytechnique de Grenoble in 1977 and 1983 respectively. Since 1979, he has been a researcher at the Centre National de la Recherche Scientifique where he is presently Directeur

de Recherche and Vice Head of the Laboratoire d'Automatique de Grenoble. He is author or co-author of over 80 journal or conference papers. His current research interests are in linear systems and adaptive control.

Luc Dugard was born some years ago in Vitry-le-François, not far from the famous Champagne vineyards, France. He received the Engineer degree in Electronics in 1975 from the Institut National Polytechnique de Grenoble. He got his "Thèse de Docteur-Ingénieur" degree in 1980 and his "Thèse de Docteur d'Etat ès Sciences" degree in 1984, both from the Institut

National Polytechnique de Grenoble.

Since 1977, he has been with the Laboratoire d'Automatique de Grenoble, E.N.S.I.E.G., where he holds a researcher position at the C.N.R.S. (the French National Center for Scientific Research). His main scientific interests are in the field of adaptive control: theoretical and methodological aspects, and applications to robotics and

thermal processes. He also has other interests, and some expertise, among others, in spoonerisms.



Benjamin Friedlander received the B.Sc. and the M.Sc. degrees in electrical engineering from the Technion-Israel Institute of Technology in 1968 and 1972, respectively, and the Ph.D. degree in electrical engineering and the M.Sc. degree in statistics from Stanford University in 1976.

From 1968 to 1972 he served in the Israel Defense Forces as an electronic engineer. From 1976 to 1985 he was at Systems Control Technology, Palo Alto, as Manager of the Advanced Technology division, where he was responsible for research and development projects in signal processing and control. During this period he was also a lecturer at Stanford University.

From November 1985 to July 1988 he was with Saxpy Computer Corporation, Sunnyvale, as the Director of Advanced Technology responsible for the research and development activities of the company. Currently he is with Signal Processing Technology in Palo Alto.

Dr Friedlander has over 240 publications in signal processing and estimation. He was an associate editor of the *IEEE Transactions on Automatic Control* in 1984, a member of the Administrative Committee of the Acoustics, Speech, and Signal Processing (ASSP) society, a member of the Technical Committee on Spectrum Estimation of the ASSP, and is the Vice Chairman of the bay area chapter of the Signal Processing society. Dr Friedlander was the recipient of the 1983 ASSP Senior Award, the 1985 Award for the Best Paper of the Year from the European Association for Signal Processing (EURASIP) and the 1989 Technical Achievement Award of the Signal Processing Society. He is a fellow of the IEEE and a member of Sigma Xi.

His current interests include advanced techniques for array processing and spectral analysis, adaptive filtering, sonar, radar and communication processing, detection tracking and localization of multiple targets, image processing, and parallel processing architectures for high speed signal processing.

Peter J. Gawthrop was born in Seascale, Cumberland, U.K. in 1952. He obtained his MA and D.Phil degrees from the University of Oxford in 1973 and 1977 respectively. He then spent some years as a Post Doctoral Research Assistant with the Department of Engineering Science at the University of Oxford, this was followed by a period as W. W. Spooner Research Fellow at New

College, Oxford. In 1981, he moved to the University of Sussex as a Lecturer, and later Reader, in Control Engineering. In 1987, Professor Gawthrop took up the Wylie chair of Mechanical Engineering at Glasgow University where he continues to teach and research in the area of Control Systems defined in its broadest sense.

His main research interests lie in the theory and application of self tuning control, system identification and system modelling. The main application areas of this research are process engineering, robotics, ships, manufacturing systems, and economic systems.

He is the author of a two volume book entitled *Continuous Time Self Tuning Control* (Research Studies Press). He is an Associate Editor of *Automatica* and is on the Editorial Board of a number of other control systems journals.

Optimization and Lagrange Multiplier Methods, (Academic Press, 1982); *Dynamic Programming: Deterministic and Stochastic Models* (Prentice-Hall, Englewood Cliffs, NJ, 1987); and co-author of *Stochastic Optimal Control: The Discrete-Time case* (Academic Press, 1978); *Data Networks* (Prentice-Hall, 1987); and *Parallel and Distributed Computation: Numerical Methods* (Prentice-Hall, 1989).



Christopher I. Byrnes was born in New York, on 28 June 1949. He received the B.S. degree from Manhattan College, Bronx, NY, in 1971 and the M.S. and Ph.D. degrees from the University of Massachusetts, Amherst, in 1973 and 1975, respectively.

He served as an Instructor in the Department of Mathematics at the University of Utah, Salt Lake City, from 1975 to 1978, when he was appointed Assistant Professor in the Department of Mathematics and in the Division of Applied Sciences at Harvard University, Cambridge, MA. From 1982 to 1985 he was Associate Professor of Applied Mathematics on the Gordon McKay Endowment at Harvard University. In 1984, he joined Arizona State University, Tempe, as a Research Professor of Engineering and Mathematics. He is currently Chairman and Professor in the Department of Systems Science and Mathematics at Washington University, St. Louis, MO, and Adjunct Professor of Mathematical System Theory at the Royal Institute of Technology (KTH), Stockholm, Sweden. Dr Byrnes has also held visiting positions at Bremen, Groningen, Harvard, IIASA, Kansas, KTH, Osaka, Paris-Dauphine, Rome-La Sapienza, Stanford and Tokyo Universities. Editor of eleven research volumes and author of over 100 technical articles, his research interests include adaptive control, algebraic system theory, distributed parameter systems, linear multivariable control, nonlinear control, and the applications of nonlinear dynamics in control and estimation. A member of AAAS, AMS, IEEE and SIAM, Dr Byrnes was named a Case Centennial Scholar by Case Western Reserve University in 1980, a Fellow of the Japan Society for the Promotion of Science in 1986 and the Graduate School Distinguished Research Professor at Arizona State University in 1988. In 1989 he was elected Fellow of the IEEE.

Dr Byrnes has served as an Associate Editor of six journals and is currently Editor of the two new book series, *Systems and Control Foundations and Applications* and *Progress in Systems and Control* published by Birkhauser, Boston, MA.



Stephen L. Campbell received the B.A. degree in mathematics from Dartmouth College, Hanover, New Hampshire, U.S.A. in 1967, and the M.S. and Ph.D. degrees in mathematics from Northwestern University, Evanston, IL, U.S.A. in 1968 and 1972, respectively. In 1972 he joined the Department of Mathematics, North Carolina State University, Raleigh, NC, U.S.A. as an

Assistant Professor, becoming an Associate Professor in 1976, and a Professor in 1981. His current research focuses on the numerical and analytical solution of implicit systems

of ordinary differential equations and their applications to control, circuit theory, and mechanics.

Dr Campbell is a member of the Society for Industrial and Applied Mathematics and its Control and Linear Algebra Activity Groups.



Ye-Hwa Chen was born in Taiwan. He received his B.S. degree in chemical engineering from the National Taiwan University in 1979, and the M.S. and Ph.D. degrees in mechanical engineering from the University of California, Berkeley, in 1983 and 1985.

From 1986 to 1988 he served as a faculty member in the Department of Mechanical and Aerospace Engineering of Syracuse University. He then became a faculty member of the School of Mechanical Engineering of Georgia Institute of Technology. His research has been on advanced control methods for manufacturing systems, adaptive robust control of uncertain systems, large-scale systems and discrete events. Dr Chen has established and served as the chairman for the Complex Systems Technical Panel of ASME since 1987. He is a member of Sigma Xi.



Emmanuel G. Collins, Jr. was born in Monrovia, Liberia, in 1959. He received the Interdisciplinary B.S. degree from Morehouse College, Atlanta, GA, in 1981, the B.M.E. degree in mechanical engineering from the Georgia Institute of Technology, Atlanta, in 1981, the M.S. degree in mechanical engineering from Purdue University, West Lafayette, IN, in 1982, and the Ph.D.

degree from the School of Aeronautics, Purdue University, West Lafayette, IN, in 1987.

In June 1987 he joined the Structural Control Group at the Government Aerospace Systems Division, Harris Corporation, Melbourne, FL. His current research interests include robust analysis and control and the use of homotopic continuation methods for reduced-order control law synthesis with an emphasis on application to large flexible space structures.



Mohammed Dahleh was born in Jordan in 1961. He received the B.S. degree in electrical engineering from Texas A & M University in 1983, and the M.A. and Ph.D. degrees from Princeton University, in 1986 and 1987 respectively.

Since 1987 he has been with Department of Electrical Engineering at Texas A & M University. His general area of interest is system and control theory; his current areas of interest include robust control, adaptive control and distributed parameter systems.



Matthew A. Dahleh was born in 1962. He received the B.S. degree in electrical engineering from Texas A & M University in 1983, and his Ph.D. degree in electrical engineering from Rice University, in 1987. He is currently an Assistant Professor of Electrical Engineering at the Massachusetts Institute of Technology. He has held consulting positions with NASA and C. S.

Draper Laboratory since 1988. His current interests include robust control, identification of uncertain systems and adaptation, control of time-varying and infinite-dimensional systems.

Dr Dahleh received the George Axelby Outstanding Paper Award with J. B. Pearson in 1989.

thermal processes. He also has other interests, and some expertise, among others, in spoonerisms.



Hüseyin Demircioğlu was born in Ankara, Turkey in 1961. He received his B.Sc. degree in electronic engineering from Hacettepe University, Ankara in 1983. From 1983 to 1985, he worked as a research and design engineer for ASELSAN (Military Electronic Industry), Ankara. He has recently obtained his Ph.D. degree from Glasgow University, U.K. His current research inter-

ests include adaptive and self-tuning control, system identification, optimal and predictive control.



Jean-Michel Dion was born in La Tronche, France, in 1950. He graduated in mathematics in 1972. He received the "Thèse de 3ème cycle" and "Thèse d'Etat" degrees both from the Institut National Polytechnique de Grenoble in 1977 and 1983 respectively. Since 1979, he has been a researcher at the Centre National de la Recherche Scientifique where he is presently Directeur

de Recherche and Vice Head of the Laboratoire d'Automatique de Grenoble. He is author or co-author of over 80 journal or conference papers. His current research interests are in linear systems and adaptive control.



Luc Dugard was born some years ago in Vitry-le-François, not far from the famous Champagne vineyards, France. He received the Engineer degree in Electronics in 1975 from the Institut National Polytechnique de Grenoble. He got his "Thèse de Docteur-Ingénieur" degree in 1980 and his "Thèse de Docteur d'Etat ès Sciences" degree in 1984, both from the Institut

National Polytechnique de Grenoble.

Since 1977, he has been with the Laboratoire d'Automatique de Grenoble, E.N.S.I.E.G., where he holds a researcher position at the C.N.R.S. (the French National Center for Scientific Research). His main scientific interests are in the field of adaptive control: theoretical and methodological aspects, and applications to robotics and

received the B.Sc. and the M.Sc. degrees in electrical engineering from the Technion-Israel Institute of Technology in 1968 and 1972 respectively and the Ph.D. degree in electrical engineering and the M.Sc. degree in statistics from Stanford University in 1976.

From 1968 to 1972 he served in the Israel Defense Forces as an electronic engineer. From 1976 to 1985 he was at Systems Control Technology, Palo Alto, as Manager of the Advanced Technology division where he was responsible for research and development projects in signal processing and control. During this period he was also a lecturer at Stanford University.

From November 1985 to July 1988 he was with Saaxy Computer Corporation, Sunnyvale, as the Director of Advanced Technology responsible for the research and development activities of the company. Currently he is with Signal Processing Technology in Palo Alto.

Dr Friedlander has over 240 publications in signal processing and estimation. He was an associate editor of the *IEEE Transactions on Automatic Control* in 1984, a member of the Administrative Committee of the Acoustics, Speech, and Signal Processing (ASSP) society, a member of the Technical Committee on Spectrum Estimation of the ASSP, and is the Vice Chairman of the bay area chapter of the Signal Processing society. Dr Friedlander was the recipient of the 1983 ASSP Senior Award, the 1985 Award for the Best Paper of the Year from the European Association for Signal Processing (EURASIP) and the 1989 Technical Achievement Award of the Signal Processing Society. He is a fellow of the IEEE and a member of Sigma Xi.

His current interests include advanced techniques for array processing and spectral analysis, adaptive filtering, sonar, radar and communication processing, detection tracking and localization of multiple targets, image processing, and parallel processing architectures for high speed signal processing.



Peter J. Gawthrop was born in Seascale, Cumberland, U.K. in 1952. He obtained his MA and D.Phil. degrees from the University of Oxford in 1973 and 1977 respectively. He then spent some years as a Post Doctoral Research Assistant with the Department of Engineering Science at the University of Oxford, this was followed by a period as W. W. Spooner Research Fellow at New

College, Oxford. In 1981, he moved to the University of Sussex as a Lecturer, and later Reader, in Control Engineering. In 1987, Professor Gawthrop took up the Wylie chair of Mechanical Engineering at Glasgow University where he continues to teach and research in the area of Control Systems defined in its broadest sense.

His main research interests lie in the theory and application of self tuning control, system identification and system modelling. The main application areas of this research are process engineering, robotics, ships, manufacturing systems, and economic systems.

He is the author of a two volume book entitled *Continuous Time Self Tuning Control* (Research Studies Press). He is an Associate Editor of *Automatica* and is on the Editorial Board of a number of other control systems journals.



Fouad Giri received his degree in electrical engineering in 1982, the Doctorat d'Etat in Automatic Control in 1988 (both from the Ecole Mohammadia d'Ingénieurs in Rabat, Morocco), and the Doctorat in Automatic Control and Signal Processing in 1988 from the Institut National Polytechnique in Grenoble (France). From 1982 to 1986, he was an assistant Professor in the Ecole

Mohammadia d'Ingénieurs. From 1986 to 1988 he was a researcher in the Laboratoire d'Automatique de Grenoble. Since 1988 he has been a Lecturer of Automatic Control at the Ecole Mohammadia d'Ingénieurs. His research interests are in adaptive and robust control.



David C. Hyland received the B.S., M.S. and Sc.D. degrees in aeronautics from the Massachusetts Institute of Technology, Cambridge, MA, in 1969, 1971 and 1973, respectively. After serving as a vibration specialist in a Cambridge-based acoustics consulting firm, in 1974 he joined the staff at Lincoln Laboratory, Massachusetts Institute of Technology. His work there included

reentry vehicle dynamics, multibody spacecraft dynamics simulation and spacecraft attitude control. In 1983 he joined the Government Aerospace Systems Division, Harris Corporation, Melbourne, FL, where he presently leads the Structural Control Group. His current research interests include robustness analysis for control-system design with application to vibration suppression in large flexible space structures.



Keith Glover was born in Bromley, Kent, U.K. in 1946. He received the B.Sc.(Eng) degree from Imperial College, London, in 1967, and the S.M., E.E. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, in 1971, 1971 and 1973, respectively, all in electrical engineering.

From 1967 to 1969 he was a development engineer with the

Marconi Company. From 1973 to 1976 he was on the faculty of the University of Southern California, Los Angeles. Since 1976 he has been with the Department of Engineering, University of Cambridge, U.K., where his present position is Professor of Engineering. He was a Kennedy Fellow at MIT from 1969-1971 and a Visiting Fellow at the Australian National University, Canberra, in 1983-1984.

His current research interests include linear systems, model approximation, robust control and identification.



Altuğ İftar was born in Istanbul, Turkey on 28 June 1960. He received his B.S. degree in electrical engineering from Boğaziçi Üniversitesi, Istanbul, Turkey, in 1982 and M.S. and Ph.D. degrees in electrical engineering from The Ohio State University, Columbus, Ohio, U.S.A., in 1984 and 1988 respectively. He held teaching assistantship positions at Boğaziçi

Üniversitesi from 1980 to 1982 and graduate teaching and research associateship positions at the Ohio State University from 1983 to 1988. He has been a research associate with the Department of Electrical Engineering at University of Toronto, Canada since September 1988. His research interests include large scale systems, decentralized control, robust control, optimization and optimal control.



Keqin Gu was born in Lanxi, Zhejiang, China, in 1957. He received his B.S. degree in 1982 and M.S. degree in 1985, both in mechanical engineering from Zhejiang University, China. He received his Ph.D. degree in mechanical engineering from the Georgia Institute of Technology, Atlanta, USA, in 1988.

From May to August 1985, he served as a member of the

Faculty of Mechanical Engineering, Zhejiang University. From September 1985 to December 1988, he worked as a Research Assistant at the Georgia Institute of Technology. From January 1989 to August 1990 he worked as a Research Associate in the Center of Robotics and Advanced Automation, School of Engineering and Computer Science, Oakland University, Michigan, U.S.A. He is currently an Assistant Professor in the Department of Mechanical Engineering, Southern Illinois University at Edwardsville, IL, U.S.A. His research interests are: robotics, adaptive and robust control, nonlinear systems theory and chaotic behaviors.



Alberto Isidori was born in Rapallo, Italy, in 1942. He graduated in electrical engineering from the University of Rome in 1965. Since 1975, he has been Professor of Automatic Control at this University. He has held visiting positions at the University of Florida, Gainesville (1974), Washington University, St. Louis (1980, 1983), the University of California, Davis (1983), Arizona

State University (1986, 1989), the University of Illinois (1987) and the University of California, Berkeley (1988). Since 1989 he is also affiliated with Washington University, St. Louis. Professor Isidori received in 1981, with coauthors, the IEEE Control Systems Society's Outstanding Paper Award. He is the author of the book *Nonlinear Control Systems* (1985, 1989). Professor Isidori is an Associate Editor of *Automatica*, *Mathematics of Control Signals and Systems* and *Applied Stochastic Models and Data Analysis*, and Associate Editor at large of the *IEEE Transactions on Automatic Control*. He is Vice Chairman of IFAC's Technical Committee on Mathematics of Control and in 1989 he was the IPC Chairman of the first IFAC Symposium on Nonlinear Control Systems Design. His research interests include systems and control theory, with emphasis on nonlinear feedback systems, geometric control theory and stabilization.



— was born in Hamburg, Germany, on 22 December 1940. He received the Dipl.-Ing. degree in electrical engineering and the Dr.-Ing. degree in guidance and control from the Technical University of Berlin, F.R.G., in 1967 and 1971, respectively. In 1980, he habilitated as privat-dozent (Dr. habil.) for teaching in the field of man-machine systems of aero-

nautics and astronautics at the Technical University, Aachen, F.R.G. From 1967 to 1971, he was a research assistant at the Institute of Aircraft Guidance, Technical University of Berlin. From 1971 to 1982, he was head of the Human Operator Branch, Research Institute for Human Engineering (EGAN/FAT), Wachtberg-Werthhoven, F.R.G., as well as lecturer of manual vehicle control, Technical University, Aachen, from 1974 to 1982. During 1977-1978, he was a visiting faculty member in the Department of Mechanical and Industrial Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, U.S.A., where he held an appointment as adjunct associate professor from 1978 to 1981. Currently, he is professor of Systems Engineering and head of the Laboratory for Man-Machine Systems, Department of Mechanical Engineering, University of Kassel, Germany. His research interests are in human control and problem solving in man-machine systems, vehicle and process control, computer-aided systems design, graphical design, human-machine dialogue systems, human-manipulator interaction, and knowledge-based systems. Dr. Johannsen is a member of several German professional societies (DGLR, VDI/VDE-GMA, VDE-ITG, GI, GfA), a senior member of the Institute of Electrical and Electronics Engineers (IEEE-SMC, Computer, CS, Robotics and Automation), a member of the Human Factors Society, and chairman of the Committee on Systems Engineering of IFAC in which he was chairman of the working group on man-machine systems



Yeo-Chow (Joe) Juan received the B.S. degree in power mechanical engineering from the National Tsing-Hua University, Taiwan, R.O.C., in 1980, M.S. degree in engineering mechanics from the University of Alabama, Tuscaloosa, in 1984, and Ph.D. degree in aerospace engineering from the University of Michigan, Ann Arbor, in 1988.

From 1986 to 1989 he worked for the Automated Analysis Corporation as a senior research engineer. He is presently a research engineer in the Engineering Technology Service of Ford Motor Company, Dearborn, Michigan. His research areas are active/adaptive control, acoustic and vibration control. Dr. Juan is a member of the American Institute of Aeronautics and Astronautics.

Pierre T. Kabamba was born in Lubumbashi, Zaire. He received the Diplome d'Ingénieur Civil en Mathématique Appliquées from the University of Louvain, Belgium, in 1977, and the Ph.D. degree in mechanical engineering from Columbia University, NY, in 1981.

He held research and teaching positions at the University of Louvain from 1977 to 1979 and

from 1981 to 1983, and at Columbia University from 1979 to 1981. Since 1983, he has been with the Department of Aerospace Engineering, the University of Michigan, where he is currently Associate Professor. His interests include dynamics, reduced-order control, model reduction, large space structures, guidance and navigation.



Dieter Kaenbauer was born in Landshut, Germany, in 1944. He received the M.S. degree in mathematics from the Technical University of Munich, in 1970, and the Dr. Techn. degree from the Technical University of Graz in 1980. Since 1971 he worked as a research scientist in the control group of the Institute for Flight Systems Dynamics at the German Aerospace Research Establish-

ment (DLR) in Oberpfaffenhofen. His research interests are in robustness analysis and control systems design.



Richard O. LaMaire was born in El Paso, Texas, on 19 October 1958. He received the B.S. degree in electrical engineering and economics in 1981 from Carnegie Mellon University, Pittsburgh, PA, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology in 1983 and 1987, respectively. At MIT,

he conducted research in the areas of adaptive and digital control theory and estimation. From 1987 until 1989, he was employed as a Member of the Technical Staff at ALPHATECH, Burlington, MA, where he worked in the areas of communications and estimation. In 1989, he joined the High Bandwidth Systems Laboratory of the IBM T. J. Watson Research Center, Yorktown Heights, NY, as a Research Staff Member. Dr. LaMaire's current interests are in the areas of communication and computer performance analysis, queueing theory, flow control in communication systems and architectures for high-speed communication systems.



David J. N. Limebeer was born in Johannesburg, South Africa, in 1952. He received the B.Sc. degree in electrical engineering from the University of the Witwatersrand in 1974, and the M.Sc. and Ph.D. degrees in 1977 and 1980, respectively, from the University of Natal in South Africa.

He was a Research Assistant at the University of Cambridge, U.K. between 1980 and 1983. In 1983 he moved to the Department of Electrical Engineering, Imperial College, London, as a lecturer. In 1989 he was promoted to Reader in Control Engineering. His research interests include multi-variable systems theory, computer aided control system design, power system stability and the control of tokomaks.





Nan K. Loh received his B.Sc. degree in electrical engineering from the National Taiwan University, Taiwan, in 1961, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Waterloo, Canada, in 1964 and 1968, respectively.

Dr Loh taught at the University of Iowa from 1968 to 1978, and joined the School of Engineering and Computer Science at Oak-

land University in 1978. Dr Loh is the John F. Dodge (Endowed Chair) Professor of Engineering at Oakland University. He is the Director of the Center for Robotics and Advanced Automation at Oakland University, a center of research excellence in Michigan which he co-founded in 1981. Dr Loh has served as a department chairman and the Associate Dean for Graduate Studies and Research, and is now the Associate Dean for Research and Development in the School of Engineering and Computer Science, Oakland University.

Dr Loh has served as an engineering consultant for numerous national and international organizations and corporations. His consulting activities include industrial development, engineering design, high technology research and development, and government cabinet level policy and goal development in science and technology. Dr Loh is a member of the board of directors of several professional organizations. He also serves on the editorial boards of several technical journals.

His areas of research interest include control systems, computer vision, robotics, digital signal processing, estimation theory and time series analysis. He has authored and co-authored over 170 technical papers and reports.



C. C. H. Ma received the BSCE degree in 1982 and the Ph.D. degree in 1986, both in electrical engineering from the University of Waterloo, Canada. He has worked briefly as a consultant, and as a staff scientist with the Institute for Computer Applications in Science and Engineering (ICASE) at NASA Langley Research Center, Virginia from 1986 to 1987. He is now an assistant

professor at the University of British Columbia, Canada. His current research interests are adaptive control, robotics, neural networks and sensor technology and their industrial applications.



John MacGregor was born in Canada in 1943. He obtained a B.Eng. degree in chemical engineering from McMaster University in 1965, and M.Sc. degrees in both chemical engineering and statistics from the University of Wisconsin in 1967. After working in the Process Technology Section of Monsanto Company in Texas between 1967 and 1969, he returned to the

University of Wisconsin to obtain his Ph.D. in statistics in 1972. He is currently a professor and chairman of the Department of Chemical Engineering at McMaster University in Canada. His research interests include statistical process control, process identification nonlinear and robust control, and the modelling and control of polymerization

reactors. He is actively involved with industries both as a consultant, and through the cooperative research efforts of the McMaster Advanced Control Consortium and the McMaster Institute for Polymer Production Technology.



John B. Moore was born in China in 1941. He received his Bachelor and Masters degrees in electrical engineering in 1963 and 1964 respectively, and his doctorate in electrical engineering from the University of Santa Clara, California, in 1967. He was appointed Senior Lecturer at the University of Newcastle in 1967, and promoted to Associate Professor in 1968 and full Professor (personal

chair) in 1973. Since 1982, he has been a Professorial Fellow in the Department of Systems Engineering, Research School of Physical Sciences, Australian National University. He has held visiting academic appointments at the University of Santa Clara (1968); the University of Maryland (1970); Colorado State University and Imperial College (1974); the University of California, Davis (1977); the University of Washington, Seattle (1981); Cambridge University and the National University of Singapore (1981); and the University of California, Berkeley (1987, 1989). He has spent periods in industry as a design engineer and as a consultant.

Dr Moore's current research is in control and communication systems. He is co-author with Brian Anderson of three books: *Linear Optimal Control* (Prentice-Hall, 1971); *Optimal Filtering* (Prentice-Hall, 1979); and *Optimal Control—Linear Quadratic Methods* (Prentice-Hall, 1989). He is a Fellow of the Australian Academy of Technological Sciences, a Fellow of the IEEE and a Fellow of the IEA Australia.



Mohammed M'Saad was born in Angads-Oujda, Morocco, in 1953. He graduated from the Ecole Mohammadia d'Ingénieurs, Rabat, Morocco, in 1978 as an electrical engineer. He obtained the degree of Docteur de 3ème cycle from the Faculté des Sciences, Rabat, Morocco, and Docteur d'Etat from the Institut National Polytechnique de Grenoble, France, in 1982 and 1987,

respectively. He was Maître-Assistant and Maître de Conférence at the Ecole Mohammadia d'Ingénieurs and researcher at the Laboratoire d'Electronique et d'Etude des Systèmes Automatiques. He is currently researcher at the Centre National de la Recherche Scientifique (C.N.R.S.), France. His research interest is adaptive control.



Rudolf Muench was born in Coburg, Germany, in 1958. He studied electrical engineering at the Technical University in Munich. For his thesis he worked at the Institute for Flight Systems Dynamics at the German Aerospace Research Establishment (DLR) in Oberpfaffenhofen and received the Dipl.-Ing. degree in 1986. He is now concerned with the development of control systems for paper machines at Sulzer Escher Wyss GmbH, Department for Research and Development, in Ravensburg.



Denis Mustafa was born in London in 1966. In 1986 he received the B.A. degree in engineering science from Oxford University, where he was a National Engineering Scholar. He received the Ph.D. degree in control engineering from Cambridge University in 1989. Since then he has been a Harkness Fellow at the Laboratory for Information and Decision Sys-

tems at the Massachusetts Institute of Technology. His research interests include robust multivariable control, model approximation, and linear system theory.



Ümit Özgüner received his Ph.D. from the University of Illinois in 1975. He has held research and teaching positions at I.B.M. T.J. Watson Research Center, University of Toronto and Istanbul Technical University. He has been with the Ohio State University since 1981 where he is presently Professor of Electrical Engineering. His areas of research interest are in decentral-

ized control in general and applied flexible structure control. He is Associate Editor in charge of book reviews for the IEEE Control Systems Transactions, is the Chairman of the IEEE Control Systems Society Working Group on Decentralized Control and a member of the Board of Governors of the IEEE Control Systems Society.



Mania Pavella was born in Nafplion, Greece. She received the electrical (electronics) engineering degree and the degree of "Docteur es Sciences Appliquées", both from the University of Liège, where she is currently Professor in the Department of Electrical Engineering. Her research interests lie in the field of electric power system analysis and control.



N. Leonard Segall was born in Montreal, Quebec, Canada, on 19 September 1957. He received his B.Sc. degree in mathematics and engineering in 1979 from Queen's University, Kingston, Ontario. From 1979 to 1981 he worked as a chemical engineer for Dow Chemical Canada in Sarnia, Ontario. He obtained his M.Eng. Degree in chemical engineering in 1983 from McMaster University

in Hamilton, Ontario. He has been working as a process control applications engineer at Esso Chemical Canada in Sarnia since 1988, while completing his Ph.D. in chemical engineering at McMaster University. His main interests are in the theory and practical application of process control.



Torsten Söderström was born in Malmö, Sweden, in 1945. He received the M.Sc. degree (civilingenjör) in engineering physics in 1969 and the Ph.D. degree in automatic control in 1973, both from the Lund Institute of Technology, Sweden. In 1976, he was awarded the title of Docent in automatic control.

In the period 1967-1974 he held various teaching positions at the Lund Institute of Technology. Since 1974, he has been working at the Department of Technology, Uppsala University, Sweden, where he has been head of the Automatic Control and Systems Analysis Group since 1975. He has worked as Lecturer and Docent and is currently Professor of automatic control.

Dr. Söderström is author or co-author of many technical papers. His main research interests are in the fields of system identification, signal processing, process control and adaptive systems. In 1981 he was, with co-authors, given an Automatica Prize Paper Award. He is a co-author of three books: *Theory and Practice of Recursive Identification* with L. Ljung (MIT Press, 1983), *Instrumental Variable Methods for System Identification* with P. Stoica (Springer, 1983), and *System Identification* with P. Stoica (Prentice Hall, 1989).



Petre Stoica was born in Rimnicu Vilcea, Romania, on 23 July 1949. He received the M.Sc. and Ph.D. degrees, both in automatic control, from the Bucharest Polytechnic Institute in 1972 and 1979, respectively. Since 1972, he has been with the Department of Automatic Control, the Polytechnic Institute of Bucharest, Romania. His research interests include various aspects of system iden-

tification, time series analysis, signal processing and adaptive control. He has authored or co-authored more than 200 papers and technical reports on the above topics and has received several awards for his publications. His most recent book (with Torsten Söderström) is *System Identification* (London, Prentice Hall, 1989).

Dr. Stoica is an Associate Editor of the *Journal of Forecasting*. He was given the Member of Time Series Analysis and Forecasting (MTSA & F) honours award.



Teng-Tsow Tsy was born in Singapore in 1961. He received the B.Eng. (electrical) degree with first class honours from the National University of Singapore in 1985 and the Ph.D. degree in systems engineering from the Australian National University in 1989. He is currently a lecturer with the Department of Electrical Engineering, National University of Singapore. His current re-

search interests include design of robust optimal controllers, robust parameters estimation techniques and adaptive control systems.



Alberto Tesi received the Laurea in electronic engineering from the Università di Firenze, Italy, in 1984. Upon completion of this degree, he worked as a system engineer at Autostrade s.p.a until 1986. In 1989 he obtained a Ph.D. in Automatic Control from the Università di Firenze. In 1990 he joined the Dipartimento di Sistemi e Informatica, where he is currently a researcher in the area

of Automatic Control and System Theory. His research interests are mainly in linear and nonlinear systems analysis, robustness and optimization.



John N. Tsitsiklis was born in Thessaloniki, Greece, in 1958. He received the B.S. degree in mathematics (1980), and the B.S. (1980), M.S. (1981) and Ph.D. (1984) degrees in electrical engineering, all from the Massachusetts Institute of Technology, Cambridge, Massachusetts.

During the academic year 1983-1984 he was an acting Assistant Professor of Electrical

Engineering at Stanford University, California. Since 1984, he has been with the Electrical Engineering and Computer Science Department at the Massachusetts Institute of Technology, where he is currently Associate Professor. His research interests are in the areas of parallel and distributed computation, systems and control theory, and applied probability.

Dr Tsitsiklis is the coauthor, with D. Bertsekas, of *Parallel and Distributed Computation: Numerical Methods* (1989). He has been a recipient of an IBM Faculty Development Award (1983), an NSF Presidential Young Investigator Award (1986), an Outstanding Paper Award from the IEEE Control Systems Society (for a paper coauthored with M. Athans, 1986), and of the Edgerton Faculty Achievement Award from M.I.T. (1989). He is an associate editor of *Applied Mathematics Letters* and the *IEEE Transactions on Automatic Control*.



Lena Valavani is currently Associate Professor in the Department of Aeronautics and Astronautics, at the Massachusetts Institute of Technology. She was formerly Boeing Assistant Professor of Aeronautics and Astronautics in the same department. Her research interests lie in the areas of robust and adaptive control, where she has published extensively and in the application

of state of the art control theory to engineering systems. Her recent pioneering work in active control of rotating stall and surge in compressors—"Smart Engines"—is a notable example. Dr Valavani is a member of the Gas Turbine Laboratory and the Laboratory for Information and Decision Systems at MIT. She has been a consultant for the C. S. Draper and Lincoln Laboratories. Dr Valavani is an Associate Editor of *Automatica*, and the newly established *Journal for Nonlinear Control and Applications* and a past Associate Editor of *IEEE Transactions on Automatic Control*. Dr Valavani is an Associate Fellow of AIAA and a member of the board of governors of the AIAA New England Section.



Vicino received the Laurea in electrical engineering from the Politecnico di Torino, Italy, in 1978. From 1979 to 1982 he held several Fellowships at the Dipartimento di Automatica e Informatica of the Politecnico di Torino. He was Researcher of Automatic Control from 1983 to 1987 at the same Department. In 1987 he joined the Dipartimento di Sistemi e Informatica,

Università di Firenze, Italy, as Associate Professor of Automatic Control. Presently, he is Professor of Automatic Control and System Theory. His research activities are mainly in the fields of robust stability and control, applied system modelling and time series prediction, robust identification. Prof. Vicino is a member of GRIS (Gruppo Riceratori Informatica e Sistemistica).



Louis Wehenkel was born in Nürnberg, Germany, in 1961. In 1986, he graduated in electrical (electronics) engineering and in 1990 he obtained his Ph.D. degree, both from the University of Liège, where he is currently a research assistant of the F.N.R.S. in the Department of Electrical Engineering. His research interests lie mainly in the field of artificial intelligence methodol-

ogies and their application to power systems



Joseph D. Wright received a B.Sc. degree in chemical engineering from the University of Alberta, Edmonton, Alberta, in 1963 and a Ph.D. degree in control engineering from Cambridge University, Cambridge, England, in 1967.

Upon graduation, he worked with Gulf Oil Canada until 1969. Joining the Department of Chemical Engineering, McMaster

University, as an Assistant Professor in 1969, he became Associate Professor in 1974 and Professor in 1979. He was visiting Professor of Chemical Engineering at the University of Alberta in 1975-76. In 1977 he became Principal Engineer, and in 1978 Manager of the Materials Processing Laboratory at the Xerox Research Centre. In 1985-86 he spent a year in the Xerox Corporation in Webster, New York, where he was Manager of the Technology Strategy Office. He returned to the Canadian Research Centre in 1986 to become Manager of the Technology and Engineering Systems Laboratory. In 1987 he was appointed Vice President and Centre Manager for the Xerox Research Centre of Canada where he leads the materials research efforts for Xerox Corporation.

Dr Wright's industrial research interests include specialty and photogenerator chemicals, novel copolymer resins and composites, polymer processing technologies, as well as paper and special coated materials. Technology transfer from fundamental research through to pilot scale process engineering is an important aspect of his work. Dr Wright's personal research interests are in the area of computer process control, in particular the application of advanced control theory to operating process units. He has published a number of papers in the areas of stochastic control, adaptive

control and multivariable distillation system and reactor control, and has presented short courses in Computer Process Control.

Dr Wright is a member of the Chemical Institute of Canada, the Canadian Society for Chemical Engineering, the American Institute of Chemical Engineers (AIChE) and the American Chemical Society. He is Chairman of the Computers and Systems Division of AIChE and Associate Editor for their newsletter, CAST Communications. He is a past president of the Canadian Industrial Computer Society and of the Sheridan Park Research Association, and is a member of the Canadian National Committee for IFAC. He is also a member of the Editorial Advisory Board for Computers and Chemical Engineering. Finally, he is a member and current Chairman of the Ontario Centre for Materials Research.



I. A. Zohdy was born in Cairo, Egypt. He received the B.A.Sc. degree from Cairo University and the M.A.Sc. and the Ph.D. degrees from the University of Waterloo, Ontario, Canada, all in electrical engineering. He was employed as a Research Assistant by the Department of Electrical Engineering, Cairo University, as a Research and Teaching Assistant at the Univer-

sity of Waterloo, Ontario, and held an NRC graduate scholarship and a Rotary International Fellowship. He was a System Engineer with Ontario Hydro, Toronto, Canada. He is Professor of Engineering at Oakland University, Rochester, Michigan. His research interests include system theory, optimization, control systems, industrial applications, imaging, and discrete event systems.

Dr Zohdy is a registered member of the Association of Professional Engineers of Ontario, a member of the Canadian Society for Professional Engineers, and a senior member of IEEE. He has been a chairman of South East Michigan IEEE.

Addendum to Lists of Reviewers for *Automatica*

The following names were inadvertently omitted from the Lists of Reviewers for *Automatica*, 1988 (Vol. 25, No. 6, 1989) and 1989 (Vol. 26, No. 6, 1990)

1988

E. H. Abed
J. Ackermann
M. Arabacioglu
K. J. Åström

A. T. Bahill
B. R. Barmish
M. Basseville
L. Berrahmoune
J. Böhm
O. H. Bosgra
P. Brunovsky

Xi-Ren Cao
N. D. Christov
K. w. T. Chu
C. K. Chui
J. D. Cobb

M. A. Da Silva
J. Dolezal
R. Doraiswami

M. Eslami

J. S. Freudenberg

B. K. Ghosh
Xing Yuan Gu

W. Jongkind

T. Kabamba
T. Kaczorek
L. Kaszkurewicz
H. Khalil
P. V. Kokotovic
W. L. de Koning
P. R. Kumar

Ching-An Lin
H. Logemann
P. B. Loh
J. Lunze

M. Mansour
I. D. Metz
P. Misra
T. Mori

A. Nagata
I. Nagy

T. O'Reilly

L. Pandolfi
M. F. Polites

M. Schlegel
J. M. Schumacher

M. G. Safonov
M. Šebek
B. Shafai
Yen-Ping Shih
W. C. Stirling

A. Tannenbaum
A. J. Telford
S. Tzafestas

M. Vidyasagar
C. de Villemagne
R. E. de Vries

K. Wei

R. K. Yedavalli

P. Zagalak
K. Zhon
K. Zhou

1989
S. Beghelli
V. Eldem
A. G. Parlos
H. Weinert
K. K. Yamamoto
R. K. Yedavalli

A Petri Net Model for Evaluation of Expert Systems in Organizations*

DIDIER M. PERDU† and ALEXANDER H. LEVIS‡

Predicate Transition Net models of the basic logic operators are used to construct dynamic models of expert systems, the latter are used to evaluate the effect of decision aids on organizational performance.

Key Words—Petri nets; expert systems; decision aids; fuzzy logic

Abstract—A new Petri Net model of symbolic computation with fuzzy logic is presented to describe the dynamics of consultant expert systems. Then a quantitative methodology is presented for assessing to what extent the measures of performance of an organization are modified when an expert system is introduced. An example problem involving a hierarchical two decision maker organization, where the expert system is used as an aid in the fusion of inconsistent information, is described. A strategy for using the expert system is compared to two other possible strategies that may be used by a decision maker responsible for this task. Measures of performance (workload, timeliness and accuracy) are evaluated for each of these strategies. The results show that the strategy involving the use of the expert system improves significantly the accuracy of the organization, but requires more time and increases the workload of the decision maker.

INTRODUCTION

ONE OF the problems in modern decision making organizations is the increase in cognitive workload of individual decision makers. This increase is attributed to higher rate of information processing that has resulted from two factors: the increase in the availability and accessibility of data and the increase in the tempo of operations. As a result, to maintain or improve performance, decision aids have been introduced or are being proposed that aim at: (a) reducing a decision maker's workload by

carrying out mundane, but time consuming tasks such as situation assessment or evaluation of alternatives, (b) augmenting the decision maker's scope by introducing additional options; and (c) reducing human error by providing guidance through a (smart) checklist.

In all cases, the introduction of decision aids has complicated the organization design problem considerably. The problem has expanded from distributing the information and decision making functions among the organization members, as is the case in traditional organization design, or between a single decision maker (DM) and a machine, as is the case in traditional man-machine work, to allocating functions to a team of humans supported by a decision aiding system.

The second aim of a decision aid introduces more flexibility in selecting a response. Unfortunately, this flexibility comes at a price: the increase in the cognitive workload of the individual DM due to the increase in uncertainty and to the introduction of a higher level cognitive task—the management of flexibility. This phenomenon, which has appeared again and again in the analysis of distributed decision making organizations, has two consequences. One is psychological and can be described by the term *meta-decision making*, while the other is organizational and can be addressed with the use of *intelligent* decision aids that can assist in managing the flexibility. The effect of meta-decisions—or, as Einhorn and Hogarth (1981) describe it “deciding how to choose”—has also been addressed by Weingaertner and Levis (1989). This paper will focus on the analysis and evaluation of one class of intelligent decision aids, those based on expert systems. The tacit assumption for these decision aids is that they can decrease the human's workload while

* Received 6 January 1990, revised 22 May 1990, received in final form 7 June 1990. The original version of this paper was presented at the 4th IFAC Symposium on Man-Machine Systems which was held in Xian, People's Republic of China during September, 1989. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Editor A. P. Sage.

† Thomson-CSF, Architecture Systemes Avancés, Boulogne-Billancourt, France.

‡ Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A. Present address: Dept. of ECE, George Mason University, 4400 University Drive, Fairfax, VA 22030-4444, U.S.A. Author to whom all correspondence should be addressed.

improving performance. The rationale is obvious: the needed functionality can be distributed between humans and intelligent machines.

In the next section, a mathematical framework is presented for modeling decision makers in organizations and decisions aids of various types for a variety of specific contexts. It is based on High Level Petri Nets, in particular, Predicate Transition Nets. Three measures of performance are used for the evaluation of the effect decision aids have: accuracy, which is a measure of the quality of the response; timeliness of the response; and cognitive workload of the individual decision maker. In the third section, an expert system model using predicate transition nets will be described. A particular example that illustrates the approach is presented in the fourth section, with the results and their interpretation appearing in the fifth.

MODELING FRAMEWORK

A restricted class of organizations will be considered. It is assumed first that the organization consists of at least two human decision makers and that it is a team. A team is defined as an organization in which the members have a common goal, have the same interests and same values, and have activities that must be coordinated so as to achieve a higher effectiveness (Grevet, 1987). It is further assumed that they are well trained for the tasks that they have to perform and that they do not learn during the execution of a particular task.

It should be possible to draw a boundary that defines what is included in the organization and what is excluded, i.e. what resides in the external environment. Tasks that the organization must perform are generated in the environment by one or more sources which may or may not be synchronized. The organization acts upon these inputs and produces a response, including the null response, that is directed to the environment. Thus, the interface between the system and the environment is composed of the sensors and the effectors.

The elements of the organization consist of the human decision makers, data bases, processors and communication systems. A decision aid is defined as any technique or procedure that restructures the methods by which problems are analyzed, alternatives developed, and decisions taken. Decision support systems, a specific form of decision aids, do not automate a specific decision making process, but must facilitate it (Keene and Scott-Morton, 1978). Decision support systems are considered here as higher level components that may consist of processors, data bases and communication systems.

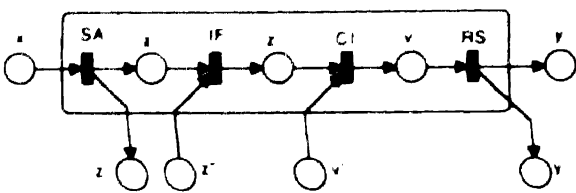


FIG. 1. Four stage model of a DM.

Relationships are the links that tie these elements together. These relationships can be considered at three levels: they may describe the physical arrangement of the components—such as the geographical location of the organization members; or the functional relationship between components—such as the sharing of information between two members; or the rules and protocols that govern the interactions—such as the conditions under which two members may share information. While this demarcation between relationships and components is often hard to justify, it is assumed that it can be done.

The Petri Net formalism (Peterson, 1981; Reisig, 1985) has been found very convenient for describing the concurrent and asynchronous characteristics of the various interactions. Petri Nets are bipartite directed multigraphs. The two types of nodes are the places, which represent signals or conditions, and the transitions, which represent processes or events. Places are denoted by circles and transitions by bars. A marking of a Petri Net assigns a non-negative integer number of tokens to each place. A transition is enabled, if and only if each of its input places contains at least one token. The places act like buffers, hosting the tokens until all the input places of a transition are non-empty. Enabled transitions can fire. When they fire, a token is removed from each input place, and a token is deposited in each output place. In the Petri Net representation of the DM model, the transitions stand for the algorithms, the connectors for the precedence relations between these algorithms, and the tokens for their input and output.

The Organization Member model

The Petri Net model of the four stage decision maker without memory who interacts with the other organization members and any decision aids present is shown in Fig. 1. The DM receives input signals x from a variety of sources: from the environment, from a decision support system (DSS), or from the rest of the organization. He can receive one input at a time at the Situation Assessment (SA) stage. He processes this input, with or without use of information stored in a data base (memory) to obtain an estimate of x .

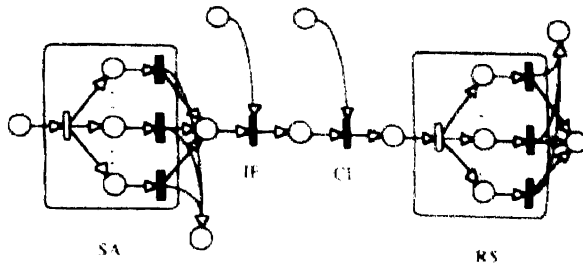


FIG. 2. Four stage model with switches.

the assessed situation z , which he may share with other DMs. He may also receive at this point other information, z'' , from the rest of the organization. He combines this information with his own assessment in the Information Fusion (IF) stage, which contains a data fusion algorithm, to obtain the final assessment of the situation, labeled z' . The next step is the consideration of commands v' from other DMs which could result in a restriction of his set of alternatives for generating the response to the given input. This is the Command Interpretation stage, or CI. The outcome of the CI stage is a signal v which contains the data z' and the rule v' used in the Response Selection (RS) stage to select the procedure or algorithm for generating the output y . This is the response of the decision maker, it may be sent to the external effectors or to other DMs within the organization.

The tokens used in Petri Net theory to study the dynamics of the nets, in the simplest version of the model, are indistinguishable. A token in a place means simply that an item of information is available to the output transition(s) of that place. It is also possible to associate attributes with the tokens. In this case, the source can be represented by a finite number of distinct tokens x , each one occurring with some probability $p(x)$. However, if the protocols ruling their processing do not vary from one set of attributes to the other, they can be considered as indistinguishable tokens.

The intelligence in this model is embodied in the algorithms embedded in the transitions; however, even if the algorithms are stochastic, the model is rather mechanistic and does not capture human decision making well. To model the choice inherent in decision making, a DM is assumed to have, at any stage of his processing, a set of options: different algorithms processing in different ways the same input to produce the same type of output. Thus, the SA and RS stages of the decision maker of Fig. 1 are modeled so as to include a set of U and V algorithms, respectively. The SA and RS stages are represented by Petri Nets with switches (Fig. 2).

Switches are transitions which resolve conflict

situations; a switch is a transition with multiple output places and a rule according to which one and only one of the output places is chosen to receive a token after the transition has fired. In the SA stage this choice is denoted by the variable u , taking its values in $\{1, 2, \dots, U\}$. The rule for determining the value of the decision variable is called the decision strategy of the decision maker for the particular stage. If the rule is characterized by a probability distribution $p(u)$ and if one branch of the switch is always chosen, i.e. if there is an i in $\{1, \dots, U\}$ such that $p(u=i) = 1$, then the strategy is called pure. Otherwise, it is mixed. The strategy that a decision maker uses at the RS stage usually depends on the input to that stage. In that case, the probabilities are conditional probabilities $p(v=j | z, v)$. Together, the strategies for the two stages constitute the internal decision strategy of the DM. While this is a way of describing the set of strategies that a well trained decision maker may use, if his bounded rationality threshold is not exceeded, there are no rules to specify how and when any of these strategies will be selected by a specific decision maker at any given time. These rules are assumed to depend on the level of expertise of the DM, and on those mental skills that "we admire but don't yet understand," i.e. on the DM's intelligence. (Minsky, 1986).

The workload of each decision maker reflects the mental effort required to carry out the information processing and the decision making. A mathematical model of workload has been developed (Boettcher and Levis, 1983) that is based on n -dimensional information theory. Its key assumption is that the higher the uncertainty in the input, the more processing has to be done to reduce uncertainty to the point that a decision can be made. The value of the workload G is obtained by computing the entropy of all the internal variables of the DM model. The cognitive limitations of human DMs can be modeled in terms of the bounded rationality constraint. This is based on the premise that the rate with which decision makers process information is bounded; if the rate is exceeded, then rapid degradation of performance occurs. Formally,

$$G/\tau \leq F_{\max}$$

where F_{\max} is the maximum rate and τ is the mean interarrival time. A recent experiment at MIT (Louvet *et al.*, 1988) has shown that for well defined cognitive tasks, F_{\max} exists, is stable, and is normally distributed across decision makers.

With this model of the organization member,

it is now possible to describe distributed decision making organizations that include decision aids. Expert Systems with their deductive capability and their ability to handle symbolic concepts have the potential to be very useful. The specific aim of this paper is to show to what extent the use of an expert system modifies the measures of performance of a decision making organization. To allow the use of the analytical framework for the study of these organizations, an expert system model using Predicate Transition Nets is first defined. Expert Systems are then studied to assess their usefulness in aiding the fusion of possibly inconsistent information coming from different sources.

AN EXPERT SYSTEM MODEL USING PREDICATE TRANSITION NETS

Knowledge Based Expert Systems show properties of synchronicity and concurrency which can be modeled by the Predicate Transition Net formalism. The rules of a knowledge base have to be checked in a specific order depending on the strategy used to solve the problem and on the current facts deduced so far by the system in the execution of previous rules. A model of an expert system using production rules to represent knowledge is presented. Some previous work (Giordona and Saitta, 1985) has addressed the modeling of production rules of a knowledge base using Predicate Transition Nets. The model presented here differs in that it incorporates explicitly the control done by the inference engine. Fuzzy logic (Zadeh, 1965, 1983; Whalen and Schott, 1983) is used to deal with uncertainty and Predicate Transition Nets are used to represent the basic fuzzy logical operators AND, OR and NOT that appear in the rules. An extension of the standard inference net formalism is obtained by the combination of these operators so that the dynamical behavior of an expert system can be represented. As a result, the rules scanned by the system to produce an answer to a specific problem can be identified and the response time, which depends on the number of rules scanned and on the number of interactions with the user, can be computed.

Predicate Transition Nets have been introduced by Genrich and Lautenbach (1981) as an extension of the ordinary Petri Nets to allow the handling of different classes of tokens. The Predicate Transition Nets used in the model have the following characteristics.

Each *token* traveling through the net has an identity and is considered to be an individual of a given class called variable. Each variable can receive different names. For this model, two

classes of tokens are differentiated. The first class, denoted by P , is the set of the real numbers between 0 and 1, representing the degrees of truth of the facts or items of evidence. The names of the individual tokens of these classes will be p , $p1$, $p2$. The second class is denoted by S . The individuals of this class can only take one value. Only one token of this class will travel through the net and will represent the action of the inference engine in triggering the different rules.

Places are nodes which can contain tokens. Three kinds of places are differentiated: (a) Those representing a fact or the result of a rule and containing either tokens of class P or no token at all; (b) Those used by the system as triggers of operators and containing the tokens of class S . These places and the connectors connected to these places are represented in bold style in the figures and constitute the *system net*; and (c) Those allowed to contain both kinds of tokens (P and S) and which are input places to transitions. The marking of a place is a formal sum of the individual tokens contained in the place. For example, a place A containing one token of class P , $p1$ and the token of class S has the marking $M(A)$:

$$M(A) = p1 + S.$$

Each *connector* has a *label* associated with it which indicates the kinds of tokens it can carry. A special grammar is used on the labels to define in what way tokens can be carried. The labels of connectors linking places to transitions contain conditions that must be fulfilled for them to carry the tokens. The labels of connectors linking transitions to places indicate what kind of token will appear in the places after the firing of the transition.

The following notation in labels is used: When token names are joined by symbol "+" then the tokens defined by these names have to be carried at the same time. For example, the label " $p + S$ " indicates that one token of the class P and one token of the class S have to be carried together at the same time by the connector. When token names are joined by the symbol ",", then the tokens defined by these names can be carried at different times but not together. For example, the label " p, S " indicates that either a token of class P or a token of class S can be carried. Mixing of notation is possible. The label " $p + S, S$ " indicates that the connector can carry either a token of class P together with a token of class S or only one token of class S . A connector without a label has no constraint on the kind of tokens it can carry.

In some cases, the connector has to carry the

token of class S when there is no token of class P involved in the firing of a transition. The statement "absence of token of the class P " is denoted by the symbol Φ . This symbol is used in the labels as if it were a class of tokens, but always in conjunction with the names of the other classes. The label " $S + \Phi$ " means that the connector can carry a token of class S , if there is no token of class P . The label " $(S + p)$, ($S + \Phi$)" means that the connector can carry either a token of class S and a token of class P or a token of class S if there is no token of class P .

Transitions have attached to them a predicate which is a logical formula (or an algorithm) built from the operations and relations on variables and tokens in the labels of the input connectors. The value (true or false) taken by the predicate of a transition depends on the tokens contained in the input places of the transition. When the predicate has the value "true", the transition is enabled. In the model of the consultant expert system, predicates are conditions on tokens of class P .

A transition without predicates is enabled as soon as all the input places contain the tokens specified by the labels of the connectors. Transitions with predicates are represented graphically with rectangles that contain the predicates. Transitions without predicates are represented with bars as in ordinary Petri Nets.

The conditions of enabling of a transition are: (1) the input places contain the combination of tokens specified by the labels of the connectors, and (2) the predicate of the transition is true. If these two conditions are fulfilled, the transition

can fire. In the firing process, tokens specified by the input connectors are removed from the corresponding input places and tokens specified by the output connectors are generated in the output places.

In order to construct the model of the expert system using Predicate Transition Nets, it is necessary to construct first models of the logical operators AND, OR, and NOT. The results are shown in Figs 3-5. Let us describe now what happens in the operator AND (the operators OR and NOT behave in a similar way). The operator drawn in Fig. 3 realizes the operation: $A \text{ AND } B \Rightarrow C$.

The operator AND can be represented as a black box, having three inputs: A , B and S_t (the trigger) and six outputs: C (on the result), A , B (memorizing of the input value) and three system places S_C , S_B and S_{next} . Only one of those system places (represented in bold style in the figures) can have a system token at the output. S_{next} will contain a system token, if the result of the operation is known, i.e. if C contains a token of the class P . This shows that the next operation can be performed. If the result is unknown, i.e. the two inputs are not sufficient to yield a result, the system token is assigned to S_C or S_B in order to obtain the values of these unknown inputs. A system token will be assigned to S_C if (i) C is unknown and (ii) A is unknown or if A and B are both unknown. The system token will be assigned to S_B if C is unknown and only B is unknown.

The execution of the operation will start only if there is a system token in S_t . We denote by S_t

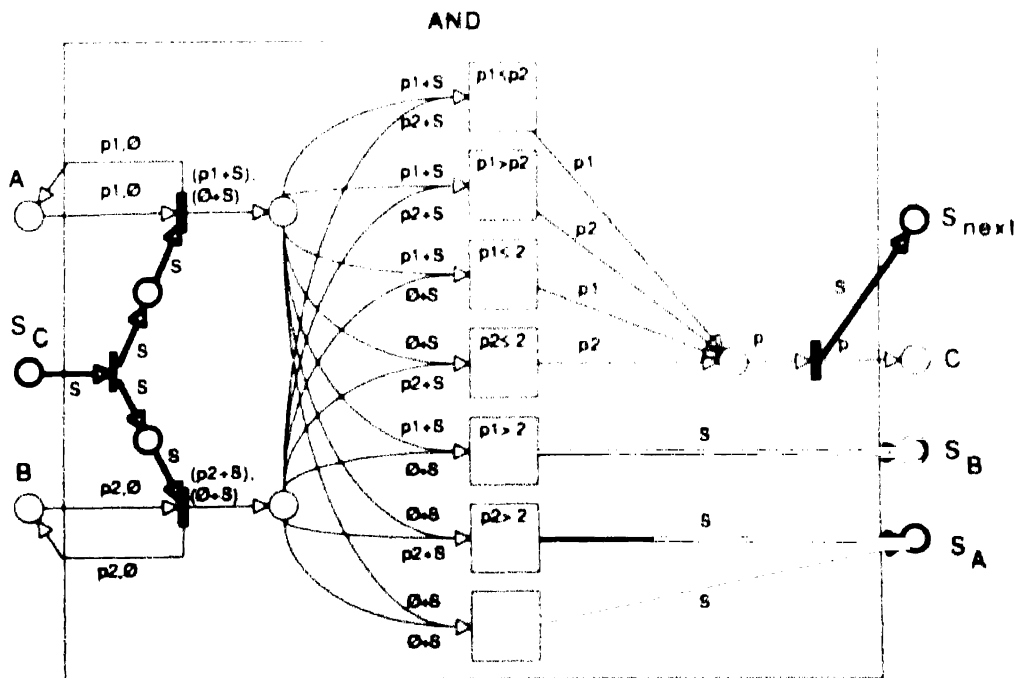


FIG. 3. Model of the operator AND

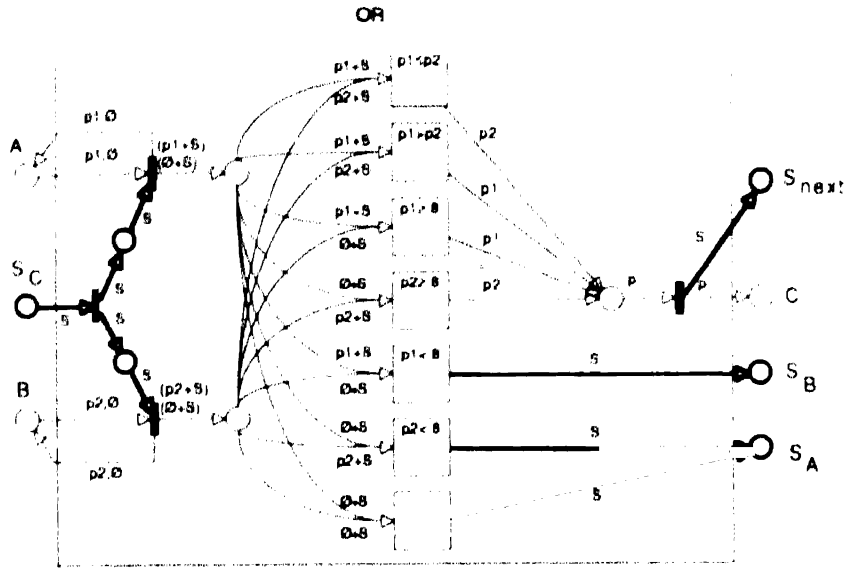


FIG. 4. Model of the operator OR

the trigger place of the operator computing C . As soon as there is a token in S_C , the two input transitions are triggered by the allocation of a system token (S) at the input places of these transitions. The values of A and B are therefore reproduced in A and B and in the output place of each of the transitions. These places contain also a system token, which will ensure the enabling of the following transition (i.e. that the two inputs are present). These two places are the input places of seven different transitions which have disjoint conditions of enabling. Only one of these transitions can be enabled and can fire. At the firing, the result, if any, is given in the result place and then in C , while the system token is assigned either to S_{next} , or to S_A , or to S_B .

These operators can be compounded in super-transitions. The model can be generalized to operators with more than two inputs by combining these basic operators.

The response time of an expert system is related to the number of rules in the rule base scanned by the system to give an answer to a specific problem or goal, and to the number of

interactions with the user. The type of model we have defined allows a quick identification of the parts of the rule base which have been scanned, given a certain set of inputs, to reach a specific goal, since each place contains the token symbolizing the value of the rule or fact it represents.

Let us consider an expert system being used to produce an answer in some environment. We represent the input X_i to the system as a n -tuple where n is the total number of questions which can be asked by the system. The answer to the questions are contained in this n -tuple at the location corresponding to the question asked (this may not be listed in order of appearance in time). The locations for the unasked questions are left empty. We denote by n_i the number of questions asked by the system. The number of X_i s might be very large but it is bounded. Given a certain environment, we can define a distribution $p_i(X_i)$ for the occurrence of the input X_i .

For a specific input X_i , we can identify N_i , the number of places scanned by the system to reach

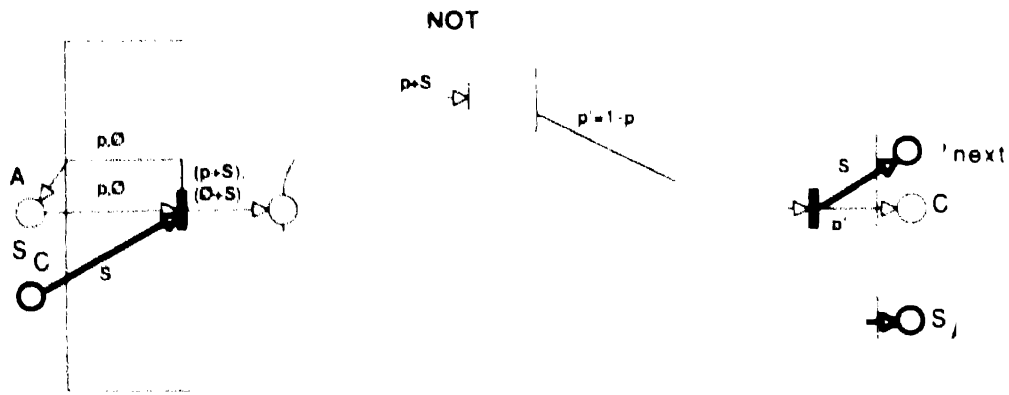


FIG. 5. Model of the operator NOT

its goal, since they all contain the degrees of truth of the subgoals they represent. If τ is the average time to check a rule and t is the average time taken by a user to answer a question asked by the system, then the time t_i to get an answer given an input X_i will be:

$$t_i = N_i\tau + n_i t.$$

Therefore, the average time of use T of the expert system for the set of inputs X_i will be given by:

$$T = E[t_i] = \sum_i p_i t_i = \sum_i p_i N_i \tau + \sum_i p_i n_i t$$

which leads to:

$$T = E[N_i]\tau + E[n_i]t$$

where $E[X]$ denotes the expected value of the variable X .

The time T obtained is the average time needed to get an answer from the expert system. This model of a consultant expert system will be used to evaluate the effect that inconsistent information can have on the command and control process.

AN EXPERT SYSTEM FOR FUSION OF INCONSISTENT INFORMATION

An important problem faced by a decision making organization is the inconsistency of information which can degrade substantially its performance. This inconsistency can be attributed to different causes: inaccuracy in measured data, lack of sensor coverage, presence of noise, bad interpretation of data. Inconsistency of information can also be explained by the attempt by a competitor or adversary to mislead about his actions through the dissemination of false information. Three strategies to fuse inconsistent information are considered: (1) ignore information sharing; (2) weighted choice among contradictory sets of data; and (3) use of an expert system which has additional knowledge on the problem to be solved.

The first strategy occurs when the decision maker performing the information fusion uses only his own assessment and ignores the assessment of the other decision maker. This strategy is related to the way a human being assigns value to information which is transmitted to him, while executing a specific task. The study of Bushnell *et al.* (1988) develops a normative-descriptive approach to quantify the process of weighting and combining information from distributed sources under uncertainty. Their experimentation has shown that one of the human cognitive biases, which appears in the execution of a task, is the undervaluing of the

communications from others, which occurs independently of the quality of the information received. The decision maker is, therefore, expected to have the tendency to overestimate his own assessment and to assign a lower value to the others' assessments.

The second strategy is to perform a weighted choice among the contradictory assessments which are transmitted to him and compared to his own. This weighting strategy involves the confidence which can be given to the information and which depends on the manner this information has been obtained, or on its certainty. In many models of organizations facing this problem of inconsistent information and using the weighted choice strategy, measures of certainty are the basis for weighting of different items of evidence. Among the methods used, the Bayesian combination has given valuable results.

The third strategy involves the use of an expert system. Expert systems can consider additional knowledge and facts which would be too costly in terms of time, effort and memory storage to be handled efficiently by the decision maker on his own. For each instance of contradictory data, it can check if their values are consistent with the knowledge it has and give an indication of their correctness. With this additional attribute, the decision maker can perform a more precise information fusion.

In order to illustrate how these strategies modify the measures of performance of an organization and to emphasize the role of an expert system in the fusion of inconsistent information, an illustrative application will be used.

The illustrative application involves an or-

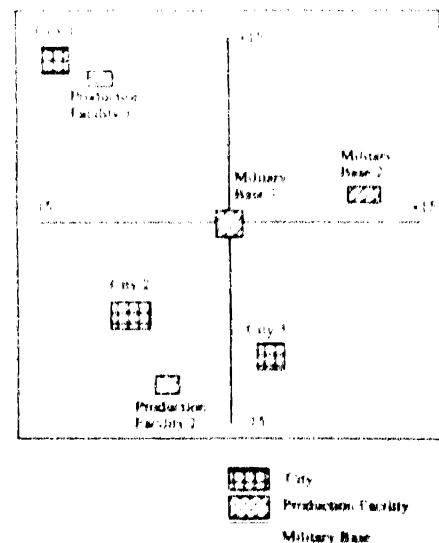


FIG. 6. Location of facilities to be defended by the organization.

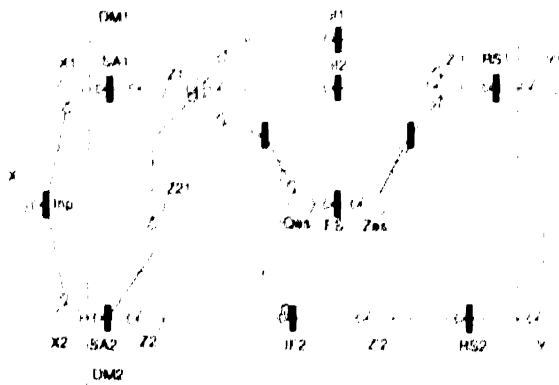


Fig. 7. Petri Net of the hierarchical 2-DM organization

ganization assigned to defend a set of facilities against air attacks. This set of facilities consists of three cities, two military bases and two production facilities located in a square, 30 miles on each side, as shown on Fig. 6. To protect itself against the incoming threats, the organization can use either one of two systems. The choice of system depends on the amount of time available, which in turn depends on the time the threat was detected and on its speed. Each system employs a different targeting solution. The performance of the organization is measured by its ability to provide the appropriate response at the right place for each incoming threat.

The Petri Net representation of the hierarchical two decision maker organization is shown in Fig. 7. The two decision makers, DM1 and DM2, perform their own situation assessment producing the results $Z1$ and $Z2$. DM2 sends $Z21$, which is equal to $Z2$, to DM1 who performs information fusion with one of the three strategies available. Using the revised situation assessment $Z'1$, the response $Y1$ is selected and transmitted to DM2 who, in turn, takes into account his new information in his information fusion stage IF2 and realizes the final response selection of the organization, Y .

Each decision maker receives as input two points on the trajectory of the threat. The first one is its position at time t , which is the same for the two decision makers to make sure they are assessing the same threat. The second point is determined by the tracking center of each decision maker. The use of decoys and the presence of noise result in the position estimates not being the same for each of the decision makers. When this is the case, we assume that one of the two is the actual position. In addition to these different coordinates, the input contains also the confidence factors associated with each position. These confidence factors have been generated by a preprocessor (say, a tracking

algorithm) and measure the quality that can be attributed to each set of data.

After receiving these inputs, the two decision makers, DM1 and DM2, perform the same situation assessment. DM1 (resp. DM2) computes the velocity of the threat and evaluates its impact point, according to the set of coordinates he has received, and produces the result $Z1$ (resp. $Z2$). DM2 sends $Z21$, which is equal to $Z2$, to DM1 who is in charge of performing the information fusion. In his information fusion stage, DM1 makes first the comparison between $Z1$ and $Z21$. If they are equal, $Z'1 = Z1$ is produced. If they are different, DM1 has to choose from the three different strategies described in the previous section.

The first one is to ignore information sharing. In this case, DM1 produces $Z'1 = Z1$ without considering the situation assessment, $Z21$, transmitted to him by DM2.

The second strategy is the weighting of the information according to the confidence factors associated with each set of data. DM1 considers the confidence levels $Conf1$ and $Conf2$ associated with the input. If $Conf1$ is greater than or equal to $Conf2$, DM1 produces $Z'1 = Z1$. In the opposite case, DM1 produces $Z'1 = Z21$.

The last strategy involves the use of an expert system. The simple knowledge base system which has been developed for this application evaluates the degree of threat as a function of the distance between the location of the different facilities and the impact point of the incoming threat as estimated by the user. A more sophisticated system could make the assessment of the threat by taking into account the type, the geographical aspect of the area, the wind direction, etc. The threat assessment is done for the two possible trajectories, one after another. If the first threat assessment shows with enough certainty that the target is one of the facilities, the computer stops its search. If not, the computer evaluates the threat if it followed the second trajectory. The answer of the expert system consists of two numbers between 0 and 1 representing the severity of the threat (according to each assessment). When the answer is given, DM1 does not use a strategy in which the result is compared with that from an internal algorithm, as described by Weingaertner and Levis (1989). This is due to the fact that the decision maker does not have enough data on his own to be able to double check the response of the decision aid. If the degree of threat according to the assessment of DM1 is greater than or equal to the one according to the assessment of DM2, the result is $Z'1 = Z1$. In the opposite case, the result is $Z'1 = Z21$.

Having chosen the trajectory which seems to be the most likely, DM1, in his response selection stage, determines the type of threat by computing the time before impact and sends it to DM2 with the fused information. DM2, in his information fusion stage, selects the system to use and performs the targeting solution in his response selection stage.

The measures of performance considered in this paper are workload (Boettcher and Levis, 1983), timeliness (Cothier and Levis, 1986) and accuracy (Andreadakis and Levis, 1987). They have been defined for the two possible types of interaction between the computer and the user.

- (a) The *user initiated mode* when the decision maker enters all the data he has in a specified order and the machine produces a result. Not all entered data may be needed by the machine in its search process.
- (b) The *computer initiated mode* when the user enters specific data only in response to requests from the computer.

Thirty-three equiprobable inputs to the organization have been considered. Twenty-four inputs contain inconsistent information. We assume that for half of these inconsistent inputs, the tracking center of DM1 is correct (the tracking center of DM2 is correct for the other half because we assume that for each input, one of the two contradictory positions is correct).

Workload. The evaluation of the workload for each decision maker uses an information theoretical framework (Boettcher and Levis, 1983). The activity of each DM is evaluated by relating, in a quantitative manner, the uncertainty in the tasks to be performed with the amount of information that must be processed to obtain certain results. The information theoretic surrogate for the cognitive workload of a decision maker is computed by adding all the entropies of all the variables used to model the procedures he uses to perform his task. The distributions of all the variables are generated by executing the algorithms for all the inputs. However, to take into account the effect of the different strategies, the workload of the decision makers has to be computed for all the mixed strategies. A mixed strategy is a convex combination of the three pure strategies.

Timeliness. The measure of timeliness considered in this application is related to the response time of the organization. A deterministic processing time has been associated with every algorithm. Again, each processing time can be described by a probability density function and the probability density function of the response time can be computed (see

Andreadakis and Levis, 1987). The use of a stochastic model does not add to the presentation of the example, but would be the model to use for an experimental investigation. For the strategy involving the use of the expert system, the time to give an answer has been computed using the expert system model described in the previous section. The response time of the expert system is a function of the number of rules scanned by the system for each input to the organization and of the number of interactions with the user. This time is likely to vary with the mode of interactions used.

We assume that DM1 and DM2 perform their situation assessment concurrently and synchronously, and that the same amount of time is needed by the two to produce an answer. Therefore, only one of the two processing times is considered. T_{SA1} , (resp. T_{RS1} , T_{SA2} , T_{IF2} and T_{RS2}) denotes the time needed to execute DM1's situation assessment algorithm (resp. DM1's response selection, DM2's situation assessment, information fusion and response selection). $T_{IF1}(i)$ is the time needed to perform the information fusion using a pure strategy i ($i = 1, 2, 3$). $T_{IF1}(3)$ is a function of the average response time of the expert system computed from its response time for all the inputs. The response time for the strategy i , $T(i)$ is therefore:

$$T(i) = T_{SA1} + T_{IF1}(i) + T_{RS1} + T_{IF2} + T_{RS2}.$$

The response time for each mixed strategy (p_1, p_2, p_3) is given by a convex weighting of the response time for each pure strategy. If $T(p_1, p_2, p_3)$ denotes the response time of the organization when the strategy (p_1, p_2, p_3) is used, we have:

$$T(p_1, p_2, p_3) = \sum p_i T(i).$$

Accuracy. The accuracy of the organization has been evaluated by comparing the actual response of the organization with the desired or optimal response expected for each input. This desired response is known to the designer. A cost of one has been attributed when the incorrect type of weapon is used or when the target point is not accurate. For each input X_i having a probability $p(X_i)$, the use of the pure strategy i generates the response Y_{ij} which is compared to the desired response Y_{di} . The cost function $C(Y_{ij}, Y_{di})$ has the following characteristics:

$$C(Y_{ij}, Y_{di}) = \begin{cases} 1 & \text{if } Y_{ij} \neq Y_{di} \\ 0 & \text{if } Y_{ij} = Y_{di} \end{cases}$$

The accuracy $J(i)$ obtained for the pure strategy

is:

$$J(i) = \sum p(X_i)C(Y_{ij}, Y_{di}).$$

The accuracy for the mixed strategy (p_1, p_2, p_3) , $J(p_1, p_2, p_3)$, obtained by computing the convex combination of the accuracy for each pure strategy:

$$J(p_1, p_2, p_3) = \sum p_i J(i).$$

Consequently, J represents the probability that an incorrect response will be generated. The lower the value of J , the better the performance is. The next section provides an analysis of the results obtained by using these measures of performance.

RESULTS AND INTERPRETATION

Using the method described above, measures of performance have been evaluated for the three strategies. For the strategy involving the use of an expert system, we have considered two different options for dealing with uncertainty in the firing of rules, Fuzzy logic or Boolean logic; and two modes of interaction between the user and the decision aid: user initiated mode or computer initiated mode. The results are summarized in Table 1.

The three first columns of Table 1 display the measures of performance (MOPs) of the organization for each pure strategy. These results show that taking into account of more knowledge, either about the way data are obtained, in the case of the weighted choice strategy, or about the meaning of the information, when the expert system is used, yields greater accuracy. Accuracy is an important measure for this kind of mission. However, the handling of more data is required. Therefore, more time is needed and more effort, expressed in terms of workload, is required. This increase in workload is caused more by the extra decisions which must be made, when the

knowledge is taken into account, than by operations or manipulation done with the additional knowledge. These manipulations are done by the decision aids, out of control of DM1.

When DM1 ignores the situation assessment of DM2, very few operations are performed. The response time is the smallest of the three. If the measure of timeliness is the ability of the organization to give a reponse as fast as possible, this strategy leads to a more timely response than the two others. The simplicity of the algorithm results in low workload for DM1 in comparison with the other strategies. This strategy has low accuracy in comparison with the other strategies, because the choice made on the information to be fused is arbitrary and has no rational justification. Thus, a clear assessment of the cost and value of coordination can be made.

For the weighted choice strategy, no operation on variables received is performed. DM1 makes only a comparison between the weights of the information. We have assumed that the weighting process was carried out outside the organization by a preprocessor and, consequently, DM1 performs only few operations more than in the first strategy. Therefore, workload and response time are slightly larger than for the first strategy because of the extra information obtained by comparing the confidence levels. An increase of 3.9% in response time and of 2.4% in the workload of DM1 is found. The use of the confidence levels, brings a large gain in accuracy: 25% improvement in comparison to the first strategy. These results show, as expected, that taking into account the quality of information plays an important role in the accuracy of the response, without degrading substantially the other measures of performance.

When the expert system is used, the increase in workload of DM1 is about 8.3% from the level of strategy 2, and 10.8% from the level of the first strategy. This can be explained by the

TABLE 1. MEASURES OF PERFORMANCE FOR THE THREE STRATEGIES

	Strategy 1 Ignoring other assessment	Strategy 2 Weighted choice	Expert system fuzzy logic		Expert system Boolean logic	
			User initiated	Computer initiated	User initiated	Computer initiated
J						
prob of error	0.360	0.270	0.210	0.210	0.240	0.240
T						
seconds	18.240	18.960	21.015	20.850	20.974	20.364
G1						
bits/symbol	63.414	64.921	70.293	70.293	65.969	65.969
G2						
bits/symbol	43.920	43.847	43.240	43.240	43.287	43.287

handling by DM1 of the assessments given by the expert system. These assessments are variables which have greater entropies and which require more processing. The increase in response time (of 10.8% from the level of strategy 2 and of 15.2% from the level of strategy 1) is mainly caused by the time taken by DM1 to interact with the system and the time needed to get the answer. This response time of the expert system can get larger as the size of the knowledge base and of the problem increase. In the example, the simplicity of the expert system hides the real effect on timeliness which can be expected with the use of such interacting system. The gain in accuracy is very significant, about 22%, when compared to the accuracy reached with the second strategy and 41.7% from the level reached when the situation assessment of the other DM is ignored. This shows the extent to which the accuracy is improved when additional knowledge is used to verify the correctness of information. By using the expert system to evaluate the threat and to estimate its severity for each possible trajectory, DM1 has a broader assessment which allows him to perform more accurate information fusion.

Finally, we note that the workload of DM2 remains almost constant for all the strategies. A variation of 1.5% can be observed. He uses always the same algorithms, and only the different distributions of the variables of the algorithms obtained, when different strategies are used by DM1, explain this small variation in his workload.

The performance measures (accuracy, timeliness, and workload of DM1) reached by the organization, when mixed strategies are used by DM1 in his information fusion stage, have been

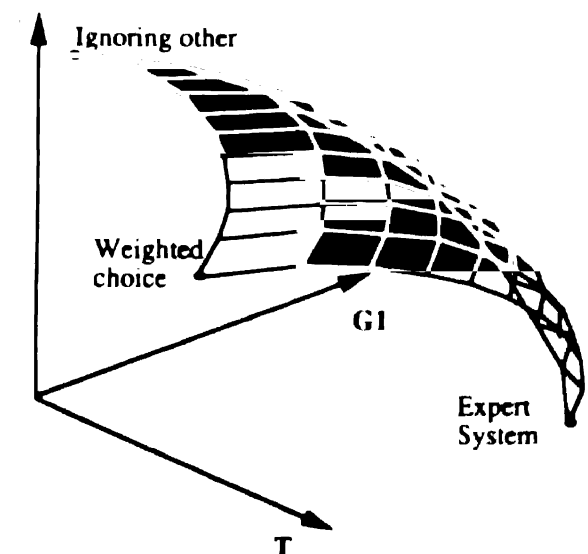


FIG. 8. Locus of the measures of performance attained by the organization.

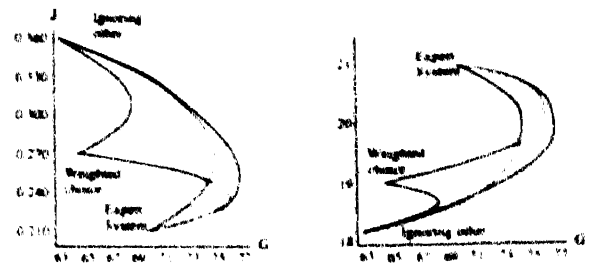


FIG. 9. Mixed strategies accuracy/timeliness to workload for DM1

obtained using CAESAR (Computer Aided Evaluation of System ARchitectures). Measures of performance have been evaluated for all mixed strategies and have led to a surface in the space (J-T-G1) represented on Fig. 8.

The projections of this surface on the Accuracy-Workload (J-G1), and Timeliness-Workload (T-G1) planes are drawn on Fig. 9. Measures of performance reached for each pure strategy are located at the three cusps of the figures. The convex combination of any two pure strategies gives a U-shaped curve (Boettcher and Levis, 1983) which can be explained by the fact that when a mixed strategy is used, there is additional activity due to the switching from one algorithm to another.

The projection of the surface of the Measures of Performance on the Accuracy-Timeliness plane (J-T) is given in Fig. 10; it shows the performance attained by the organization. The corners of this triangle indicate the level reached in accuracy and response time for each pure strategy. For all binary variations between pure strategies or for all successive binary combinations of mixed strategies, J and T are linear combinations of each other. Figure 10 shows clearly the tradeoffs between response time and accuracy and how the requirements of the mission will justify a strategy. Thus, if the requirements in accuracy are too binding, the strategy of ignoring information sharing will not be acceptable. In the same way, if the time available to process each input is too short, the expert system would be useless because too

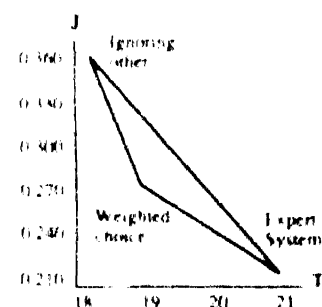


FIG. 10. Mixed strategies accuracy and timeliness of the organization

much time will be needed to perform the information fusion.

Effect of the mode of interaction

The effect of the mode of interaction on the measures of performance is shown on the last four columns of Table 1. There is no change in accuracy or workload; however, a slight change in timeliness is observed. This is caused by the fact that, in the user initiated mode of interaction, all the data which have a chance to be processed by the expert system are entered at the beginning of the session. In the example, the position of the impact points according to the different situation assessments are entered, even if the first set is sufficient to assess the threat. Therefore, more time is needed than in the computer initiated mode, where data are entered at the request of the system during the search.

It is important to note that in the air defense example, no workload has been assigned to the process of entering the information in the expert system. The process consists only of replication of the information the decision maker already has. If the inputs asked by the expert system do not correspond to the data the decision maker has, he would have to perform some operations to deduce these inputs from the information he has. Let us consider an example where the decision maker has computed or received from another member of the organization the value of the speed of an object being analyzed. If the expert system asks the decision maker the question: "speed of the object: [possible answer: low, moderate, high]," the decision maker will have to deduce from the actual value of the speed the attribute asked by the system. A small algorithm will have to be executed, increasing his workload. It can be expected therefore that, in this case, a change in workload similar to the change in response time would be observed. This issue raises the problem of the adequate design of the expert system, or more generally, of the decision aid in which the mode of interaction has to be thought very carefully to avoid an unnecessary increase in the workload of the decision maker and in the response time.

Fuzzy logic vs Boolean logic

For this illustrative application, the levels of performance reached, when different expert systems are used, have been studied. The performance achieved with an expert system using fuzzy logic as the means of inference has been compared to the performance obtained by using an expert system which does not deal with uncertainty and uses Boolean logic. This version of the expert system has been obtained by

changing the mapping functions (only values 0 and 1 could be processed instead of the real numbers between 0 and 1). It has been assumed that a statement having a degree of truth greater (resp. smaller) than 0.6 was true (resp. false). Therefore, the assessment of the threat for each trajectory has only the values true or false. The different measures of performance obtained for the two systems are summarized in the last four columns of Table 1.

The organization has a response time slightly lower with an expert system using Boolean logic than with the expert system using fuzzy logic (2.3%). This is due to the fact that by assigning the value true or false to the severity of threat, the system can reach a conclusion (which is not always the best one) by examining fewer possibilities. It can prune a larger part of the knowledge base than the fuzzy logic system when it reaches the conclusion that a specific facility is threatened. When this conclusion is reached for the the first possible trajectory, the other trajectory is not examined. This results in a shorter time to produce the answer and in fewer interactions with the user and therefore in a shorter response time.

Since the expert system with Boolean logic assesses the threat only with the value true or false, the answer of the expert system has a lower entropy. The workload of the decision maker is therefore lower (about 6.8%) when he uses the expert system with Boolean logic than when he uses the expert system with fuzzy logic.

By pruning a larger part of the knowledge base when it reaches a conclusion, the system has more chance to make the wrong assessment of the threat. The results show that, indeed, the system with Boolean logic exhibits lower accuracy than the system with the fuzzy logic. The level of accuracy is, nevertheless, better than for the two other strategies expected to be used in the information fusion stage and is explained by the fact that more knowledge is taken into account in the information fusion process.

CONCLUSION

A methodology for the modeling and evaluation of decision aids in an organizational context has been presented. Petri Nets models of human decision makers and decision aids are described that can be interconnected to represent organizations. A procedure has been presented for assessing to what extent the measures of performance of an organization are modified when a decision aid is introduced. First, a model of symbolic computation with fuzzy logic, using

Predicate Transition Nets, is presented to describe the most common kind of expert system: the consultant expert system. An illustrative example has been used to evaluate alternative strategies for handling inconsistent information. The results show that the strategy involving the use of the expert system improves significantly the accuracy of the organization, but requires more time and increases the workload of the decision maker using it.

Acknowledgement—This work was conducted at the MIT Laboratory for Information and Decision Systems with support provided by the Basic Research Group of the Joint Directors of Laboratories through the Office of Naval Research under contract no. N00014-85-K-0782.

REFERENCES

- Andreadakis, S. K. and A. H. Levis (1987) Accuracy and timeliness in decision making organizations. *Proc. 10th IFAC World Congress*, Pergamon, Oxford.
- Boettcher, K. L. and A. H. Levis (1983) On modeling teams of interacting decision makers with bounded rationality. *Automatica*, **19**, 703–709.
- Bushnell, L. G., D. Serfaty and D. L. Kleinmann (1988) Team information processing: A normative-descriptive approach. In S. E. Johnson and A. H. Levis (Eds) *Science of Command and Control: Coping with Uncertainty*. AFCEA International, Fairfax, VA.
- Cothier, P. H. and A. H. Levis (1986) Timeliness and measures of effectiveness in command and control. *IEEE Trans. Syst. Man Cybern.* **SMC-16**, 844–853.
- Einhorn, H. J. and R. M. Hogarth (1981) Behavioral decision theory: Processes of judgment and choice. *Annu. Rev. Psychol.* **32**, 53–88.
- Genrich, H. J. and K. Lautenbach (1981) System modeling with high level petri nets. *Theoret. Computer Sci.* **13**, 109–136.
- Giordana, A. and L. Sant'Anna (1985) Modeling production rules by means of predicate transition networks. *Inform. Sci.* **35**, 1–41.
- Grevet, J. L. (1987) Decision aiding and coordination in decision making organizations. S.M. Thesis, Report LIDS TH-1737, Laboratory for Information and Decision Systems, MIT, Cambridge, MA.
- Keene, P. G. W. and M. S. Scott-Morton (1978) *Decision Support Systems: An Organizational Perspective*. Addison-Wesley, Reading, MA.
- Louvet, A. C., J. L. Casey and A. H. Levis (1988) Experimental investigation of the bounded rationality constraint. In S. E. Johnson and A. H. Levis (Eds) *Science of Command and Control*. AFCEA International, Fairfax, VA.
- Minsky, M. L. (1986) *The Society of Mind*. Simon and Schuster, New York.
- Peterson, J. L. (1981) *Petri Net Theory and the Modeling of Systems*. Prentice Hall, Englewood Cliffs, NJ.
- Reisig, W. (1985) *Petri Nets: An Introduction*. Springer, Berlin.
- Wingardner, S. J. and A. H. Levis (1989) Evaluation of decision aiding in submarine emergency decision making. *Automatica*, **25**, 349–358.
- Whalen, T. and B. Schott (1983) Issues in fuzzy production systems. *Int. J. Man-Machine Studies* **19**, 57–71.
- Zadeh, L. A. (1965) Fuzzy sets. *Inform. Control*, **8**, 338–353.
- Zadeh, L. A. (1983) The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Syst.* **11**, 199–227.

A New Technique for Fault Detection Using Petri Nets*

J. PROCK†

Modelling a system as a Petri net, abnormal systems' behavior or sensor errors with slow time constants can be detected monitoring measurement signals which are related to conservation quantities.

Key Words—Failure detection; leak detection; nuclear plants; on line operation; Petri nets; sensor failures; signal processing

Abstract—A new method of fault detection using dynamic measurement signals which is suitable for larger systems is presented. By modelling the system as a Petri net, failures with very slow time constants are detectable. The method is limited to the identification of sensor or process errors which are manifested in signals related to physical conservation quantities. After a fault is detected a prognosis of the future system's behavior can be provided. The method is applied to a nuclear power plant secondary cooling loop. The advantages and drawbacks are discussed in detail.

1. INTRODUCTION

HIGHER DEMANDS upon process safety require an increased effort with respect to process control, which includes the early detection of faults and abnormal process behavior. Furthermore, decreasing hardware costs support the development of new, computer-based failure detection and identification methods which process measurement signals and work in the time range. Up to the present, faults have been detected often by applying the techniques of analytical redundancy [reviewed in Basseville and Benveniste (1986)]. These techniques are appropriate to "partial" processes. A partial process is defined hereafter as a structural component, which transports or transforms a physical conservation quantity. This definition comes from a power plant point of view.

The methods of analytical redundancy are also useful in the domain of nuclear power plants (Prock, 1989a), that is, for the online real time identification of fast changes in sensor signals during steady state or transient plant operating conditions. But the main drawback of these methods is that they cannot detect faults with

slow time constants because of their limited time memory. This means that the occurrence of a fault will be forgotten by the algorithm after a certain period of time. Faults with a low time constant will change the measurement signals during this time period only to a minimal extent and will be therefore undetectable. The limited memory span is a consequence of the principle of these methods. They need state space models of the partial process under consideration, e.g. Prock (1988), which are driven by input signals. This form of description as a boundary value problem will never be appropriate for detecting slow varying failures under real time conditions, as shown below. Therefore additional methods for early fault detection are required, as the occurrence of slow varying failures has a nonzero probability. Examples of such failures are small leaks in vessels or temperature drifts in sensors.

For the case of "total" processes which are composited by several partial processes, like bigger systems in a power plant or in a chemical factory, a new technique of online fault detection in real time is presented in a process independent formulation using place/transition nets (Section 3). Place/transition nets are a subclass of Petri nets. For readers who are unfamiliar with the Petri net theory, some of the technical terms required to derive the method are presented in Section 2. In Section 4 this new fault detection method is applied to a secondary cooling loop of a pressurized water reactor and tested with the help of original sensor data. Discussion and further prospects follow in Section 5 and the paper concludes with a brief review in Section 6.

2. BASIC DEFINITIONS

At the beginning some basic definitions of place/transition nets (hereafter called pt nets)

* Received 21 October 1989; revised 10 June 1990; received in final form 2 July 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor M. G. Robb under the direction of Editor H. Austin Spany III.

† Gesellschaft für Reaktorsicherheit, D-8046 Garching, Germany.

must be given. For a more detailed study of the theoretical aspects of Petri nets, see Reisig (1982); a description from a more practical point of view can be found in Abel (1987).

A pt net is a 6-tuple $\tilde{N} = (S, T, F, K, W, M(0))$ where

- (i) S is a finite set of places $S = \{s_1, s_2, \dots, s_v\}$;
- (ii) T is a finite set of transitions $T = \{t_1, t_2, \dots, t_r\}$;
- (iii) F is a binary relation $F \subseteq (S \times T) \cup (T \times S)$ which is represented by directed arcs between the places and transitions;
- (iv) K is a mapping, $K: S \rightarrow \mathbb{N} \cup \{\omega\}$ which describes the capacity of each place $K(s)$;
- (v) W is a mapping, $W: F \rightarrow \mathbb{N} \setminus \{0\}$ which attaches a weight to each arc; and
- (vi) $M(0)$ is a mapping, $M(0): S \rightarrow \mathbb{N}$ which gives the initial marking of the places, taking the capacity $K(s)$ of each place into account.

A pt net is a bipartite graph consisting of 2 types of elements, a finite number $/S/$ of places [definition (i)] and a number $/T/$ of transitions (ii). Definition (iii) states that these places and transitions are coupled alternatingly by arcs, which are directed from a place to a transition or *vice versa*. Pt nets are introduced to describe the transport of a quantity, such as mass, information or a sort of goods. To remain in a system independent formulation, this transport quantity will be called a token. Places, the passive

elements of the net, store a certain number (finite or infinite) of tokens according to their capacity K [definition (iv)]. Under special conditions a transition transports a number of tokens from the previous to the next place(s), the transition fires. The number of tokens transferred depends on the transport capacity [the weight W , definition (v)] of the connecting arcs. The transport process starts at time 0 with an initial distribution $M(0)$ of the tokens per place (vi). Figure 1 is an example of the graphical representation of a pt net; the places are identified by circles, and the transitions by boxes.

The static structure of the pt net, defined in (i), (ii), (iii) and (v) can also be described in an algebraic manner by the so-called $/S/$ -row $/T/$ -column incidence matrix N . The element N_{ij} of this matrix indicates if place s_i is reached by a weighted arc coming from transition t_j (+ sign) or is left by an arc which is directed to transition t_j (-). Assuming that no arc simultaneously starts and ends at the same place (the noloop condition: $(s, t) \in F \Rightarrow (t, s) \notin F$) N_{ij} is defined:

$$N_{ij} = \begin{cases} +W(t_j, s_i), & \text{if } (t_j, s_i) \in F \\ -W(s_i, t_j), & \text{if } (s_i, t_j) \in F \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The dynamic behavior of the pt net is represented by the firing rule. A transition t_i will be able to fire, if the following relation holds, where $\mathbf{M}(k)$ is a marking vector, k is a discrete

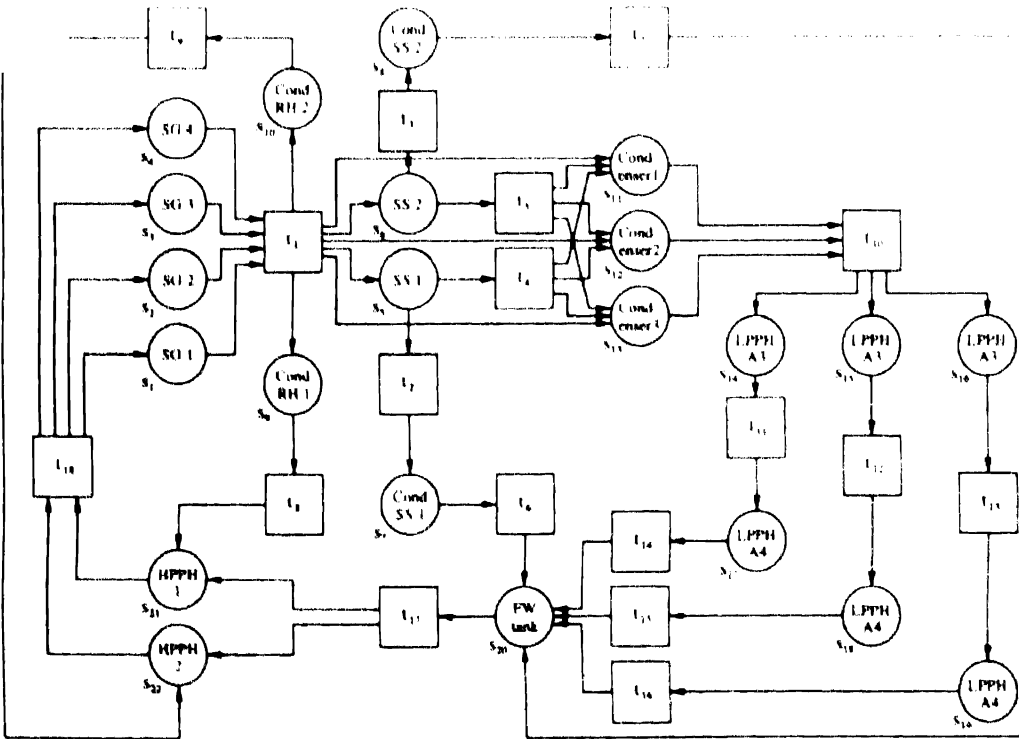


FIG. 1. Modelling of a PWR secondary cooling loop as a mass token place/transition net.

time point $(\mathbf{M}(k): S \rightarrow \mathbb{N}_0)$

$$0 \leq \mathbf{M}(k-1) + \mathbf{t}_j \leq \mathbf{K} \quad (2)$$

This means that firing will be only possible if the token content per place after firing does not surpass its capacity. \mathbf{K} is the capacity vector of dimension $|S|$ and the transition \mathbf{t}_j is the j th column vector of the incidence matrix \mathbf{N} . After firing the subsequent marking $\mathbf{M}(k)$ arises

$$\mathbf{M}(k) = \mathbf{M}(k-1) + \mathbf{t}_j \quad (3)$$

With the help of equation (3), and beginning at time $k = 0$, it follows:

$$\mathbf{M}(k) = \mathbf{M}(0) + \mathbf{N}\mathbf{v} \quad (4)$$

where vector \mathbf{v} describes the firing frequencies of each transition which lead from the initial marking to the actual state. Of particular importance are special sets of places of the pt net called S -invariants. They are integer solutions of the linear equation

$$\mathbf{i}^T \mathbf{N} = 0 \quad (5)$$

With the help of these S -invariants a new principle for fault detection in complex systems can be formulated.

3. PETRI NETS FOR FAULT DETECTION IN LARGER SYSTEMS

Petri nets are a powerful tool for system description. Nevertheless up to the present they have mainly been used for simulation purposes only. The problem of process monitoring in a power plant can be stated as follows. The measurement signals come from the system with a constant scanning rate. When processing these data, a computer-based system should decide online in real time if an error has occurred or not. To perform this, the computer program needs some knowledge about the system (or the "total" process, as described in Section 1) under consideration.

Supposing it is possible to map the structure of the total process as a pt net, the transport of the physical conservation quantity is represented by the firing of tokens. If the conservation quantity takes only a few discrete values and the signals measuring the number of tokens are not noisy, the process monitoring is easy: Using equation (4) it can be tested at each scanning time point if the actual marking vector $\mathbf{M}(k)$ beginning from the initial marking $\mathbf{M}(0)$ is reachable. If $\mathbf{M}(k)$ is not reachable it can be concluded that an error has occurred. The algorithmical evaluation of this failure detection criterion is simple. Keeping in mind that the marking vector is integer valued, equation (4) is therefore a linear diophantic equation system. It is sufficient to test

the existence condition of (4) at each time step. More details of the theory of diophantic equations can be found in Pascoletti (1986). Examples for systems governed by such well defined unnoisy physical quantities are industrial production systems or automatic shunting yards.

But in the case of plant monitoring, the measurement signals are noisy and their domain of definition is much larger than in the former case. Because of this the simple evaluation of equation (4) must fail.

Multiplying equation (4) with the transpose of the S -invariant and taking (5) into account the following equation holds

$$\mathbf{i}^T \mathbf{M}(k) = \mathbf{i}^T \mathbf{M}(0) \quad (6)$$

For the application in power plants it is correct to assume that the net token flow across the envelope surface of the total process under consideration vanishes or is zero in the mean. Otherwise continuous plant operation is not possible. Moreover it is assumed that the transitions fire without changing the number of tokens, in other words, the sum of arc-weights in front and after a transition should be equal

$$\sum W(x, t) = \sum W(t, y) \quad (7)$$

with ${}^*t := \{y \in S : (y, t) \in F\}$, $t^* := \{x \in S : (t, x) \in F\}$. Equation (7) is a conservation law for the firing of tokens. Under both these conditions it is clear that an S -invariant exists which does not contain any other elements than 1 because each column sum vanishes. Such an S -invariant is called hereafter a covering S -invariant. Therefore equation (6) can be rearranged as

$$\sum M_i(k) - \sum M_i(0) = 0 \quad (8)$$

The second sum in equation (8) must be calculated only once at the initial time $k = 0$.

Taking the noisy nature of the measurement values into consideration, a new fault criterion for continuous total processes can be formulated

$$\sum |M_i(k) - M_i(0)| < \epsilon \quad (9)$$

Equation (9) is well suited for online process monitoring. The actual number of tokens per place $M_i(k)$ is compared with the initial token content of the total process. This is possible because the continuous total process is naturally an initial boundary problem in contrast to the partial processes of the analytical redundancy methods. Therefore each slow varying fault can be detected as soon as it surpasses the threshold ϵ . The height of ϵ depends on the sensor noise

and can easily be determined in an initial learning period.

If equation (9) does not hold one of the following reasons are at cause:

- (i) A sensor fault has occurred, and one of the measured token numbers is erroneous;
- (ii) Inside the total system a source or sink of tokens has arisen, which means the structure of the pt net has changed; or
- (iii) The net token flow across the envelope surface of the total process is no longer zero mean, and the operation of the plant has become discontinuous.

Which exactly of these different faults has occurred cannot be recognized by using equation (9), that is, a diagnosis is not possible. For this to be done more knowledge about the total process under consideration in the form of quantitative or qualitative physical models is needed, but the discussion of this domain is not the scope of this paper.

At the end of this section it should be noted that the process description in terms of pt nets is analogous to the state space formulation

$$\mathbf{M}(k) = \mathbf{A}\mathbf{M}(k-1) + \mathbf{B}\mathbf{u}(k-1). \quad (10)$$

With the assumption that the state \mathbf{M} is totally measureable and that it only represents physical quantities of the same kind (i.e. only masses or only temperatures) the state transition matrix \mathbf{A} is equal to the identity matrix. Eliminating all previous time points $k-1, k-2, \dots, 1$ in equation (10) one gets

$$\mathbf{M}(k) = \mathbf{M}(0) + \mathbf{B} \sum_{j=0}^{k-1} \mathbf{u}(j). \quad (11)$$

This condition of observability (11) will be equivalent to the condition of reachability (4) if the sum of the control vector in (11) is set to a vector \mathbf{v} and the input matrix \mathbf{B} is named by incidence matrix \mathbf{N} . The analogy demonstrates that it is suitable to use the Petri net description to problems of process monitoring.

4 APPLICATION TO A PWR SECONDARY LOOP

In Fig. 1 the mass transport in a secondary cooling loop of a pressurized water reactor (PWR) is modelled by a pt net. The plant concerned is the Biblis-B nuclear power plant, Germany. To guarantee clarity, the net model is simplified with respect to the transitions. The neglecting of transitions is possible because the failure detection criterion (9) needs only the definition of places. The formulation of the incidence matrix which contains all transitions is not necessary. Transitions are represented by

boxes (containing the letter t with subscript) and places by circles (inside containing an abbreviation of the physical meaning, outside an s with subscript). A guideline for the distribution of the plant components (the partial processes) into places and transitions is given below. It should be noted that the structure of the net and the physical structure of the pipe system must differ. The net considers only the mass transport; the pipe system is designed for the simultaneous transport of mass, energy and momentum.

It is not the concern of this paper to discuss the modelling process and the physics of a secondary loop in detail. However it should be mentioned that Fig. 1 does not contain all secondary loop systems. Some processes which are to be modelled as transitions, like the extraction steam lines of the turbines or the recirculation units of condensate in the low and high pressure preheaters are not considered. Additionally, the auxiliary systems like the steam generator blowdown device and the system for the demineralization of water have also not been taken into account. The secondary loop is a two phase system. It contains steam localized to the range between the steam generators and the turbines, and water localized to the remaining components. Because of the fact that the specific density of steam is much lower than the density of water, all steam transporting components have been modelled as transitions.

For the evaluation of the token numbers per place, level measurement values must be used. With the help of the geometry of the components considered and the (temperature dependent) mass density, the total water mass per place can be calculated. In the case of the 4 steam generators the enthalpy rise values are also needed to estimate the internal, unmeasurable riser level. Table 1 lists the 22 places of the pt net of Fig. 1 and specifies the measurement signals. The code of the signals refers to the measuring instruments of the power plant as standardized by the identification system AKZ of the German Kraftwerk Union.

Knowing the token number per place the token sum $\Sigma(k)$ is calculated at each scanning time point k (in the case of Biblis-B the scanning rate of analogous signals is 1s). For the interpretation of equation (9) a low-pass filter

$$\Sigma_f(k) = \Sigma_f(k-1) + \alpha[\Sigma(k) - \Sigma_f(k-1)] \quad (12)$$

is needed, where the amplification factor α should be chosen in the order of some 10^{-2} . The use of such a filter is appropriate because of the noisy nature of Σ . This noise is a consequence of the measurement signal noise, the simplifications

TABLE 1. THE PHYSICAL MEANING OF THE PLACES OF A PWR SECONDARY LOOP PLACE-TRANSITION NET AND THE CODE OF THE SENSOR SIGNALS FOR THE TOKEN DETERMINATION

Place	Name of the partial process	Sensor signal (AKZ-code)
s_1	steam generator 1	20 YB 01 1.951 20 YA 01 1.951
s_2	steam generator 2	20 YB 02 1.951 20 YA 02 1.951
s_3	steam generator 3	20 YB 03 1.951 20 YA 03 1.951
s_4	steam generator 4	20 YB 04 1.951 20 YA 04 1.951
s_5	steam separator 1	20 RN 10 1.001
s_6	steam separator 2	20 RN 20 1.001
s_7	condensate vessel	20 RN 10 1.002
s_8	steam separator 1, 2	20 RN 20 1.002
s_9	condensate vessel	20 RP 11 1.001
s_{10}	reheater 1, 2	20 RP 21 1.001
s_{11}	condensate collecting container	20 SD 11 1.011
s_{12}		20 SD 12 1.011
s_{13}	condenser 1, 2, 3	20 SD 13 1.011
s_{14}	low pressure preheater A3	20 RH 23 1.004
s_{15}		20 RH 33 1.004
s_{16}	1, 2, 3	20 RH 43 1.004
s_{17}	low pressure preheater A4	20 RH 24 1.003
s_{18}		20 RH 34 1.003
s_{19}	1, 2, 3	20 RH 44 1.003
s_{20}	feedwater tank	20 RE 50 1.002
s_{21}	high pressure preheater 1, 2	20 RE 61 1.003
s_{22}		20 RE 62 1.003

of the modelling process and the dynamic behavior of the transient. For the online fault detection the mean of the low-pass innovation is evaluated, the length w of the moving averaging window was fixed to 100 time steps. When choosing w one must come to a compromise because the lowering of w leads to a faster detection, whereas a higher value of w increases the sensitivity of the detection. Alarm will be given if the failure sensitive signal (also called residual) surpasses a threshold ξ :

$$\sum [\Sigma(r) - \Sigma_r(j-1)]^2 \quad (13)$$

An example for such a residual is given in Fig. 2. On top it shows the residual in the nofail case and the definition of the alarm threshold derived from its nofail behavior. For real applications the principle of self-adapting, learning thresholds (Prock, 1989b), should be used. In the lower part of Fig. 2, a very small leak in the feedwater tank was simulated beginning at time 5000s. The residual indicates this fault, and reaches the threshold after 394 time steps. At time 5394 the alarm can be given. During the time range between the beginning of the fault and its detection, the water level in the

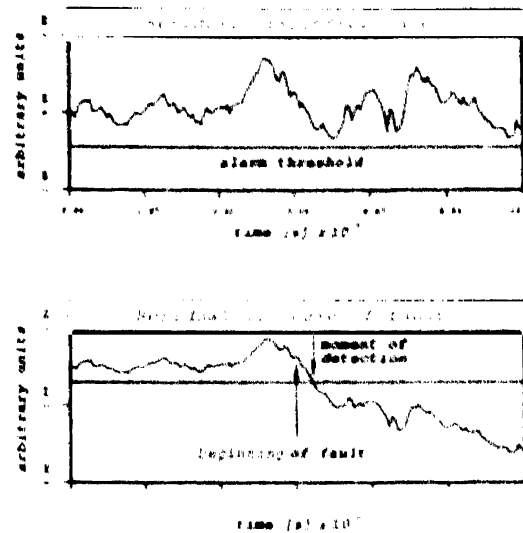


Fig. 2. Detection of a leak in the feedwater tank. Beginning of the simulated fault at time 5000s with a mass loss of 10 kg/s which is equivalent to a level change of 60 $\mu\text{m/s}$.

feedwater tank decreases by 2.4 cm which is equivalent to a total mass loss of 4000 kg. This loss is very small compared to the whole mass content of the secondary loop which shows the sensitivity of the fault detection method. At present the feedwater tank level has to decrease, in the worst case, by 55 cm before a warning is given to the control room operator. It should be mentioned that the measurement signals come from an operational transient where the generator power drops from about 60% to about 8%. If the sensor data come from a plant in a stationary state (and this is the normal operation of a nuclear power plant) it will be less dynamic and the fault will be detected earlier than in the example above.

There are other possible faults in the PWR secondary loop which are detectable with the alarm criterion (9) or (13) respectively. Today the detection of small leaks in the live steam pipes is difficult because the warning signal, which indicates the pressure gradient, operates only for leaks with a diameter larger than 30 cm. A leak or break of one steam generator U-tube is identified in existing plants by measuring the activity. Using a secondary loop Petri net model, redundant information can be provided to validate this activity signal. The same holds in the case of leaks or breaks of the condenser pipes which should be detected by measuring the electrical conductivity.

It can be summarized that the method presented is not able to localize the point of fault and is not suitable for the detection of large faults, like big leaks or the total break of pipes with a large cross section area. In these cases the calculation of the residual mean takes too much time. However, this is not a real drawback

because fast transients as a consequence of serious faults are well managed by the automatic plant safety systems. The advantage with respect to the secondary PWR loop of the method presented is the early detection of faults with a low or a very low time constant. This early detection (before an automatic safety system starts to operate) can assist in lowering the consequences of the fault.

5. DISCUSSION

There is additional information needed amongst the measurement values for the computer based detection of faults or abnormal system operation. The methods of analytical redundancy need this system describing knowledge in the form of state space models. These are processed inside the validation algorithms. Our new derived fault detection criterion based on Petri nets is simple in contrast to the validation algorithms and much less computing time is needed for its online calculation. However there is a catch; much more knowledge is needed for the modelling of the total system as a pt net. In contrast to the evaluation of the state space models of the analytical redundancy, in this case more in depth knowledge about the behavior of the total system is necessary. And so, the advantages in computer processing have to be paid at the expense of the modelling work.

The description of a (complex) total system with the help of a Petri net is not unique. Therefore only some hints can be given for the modelling work: Since the fault criterion equation (9) is a conservation principle only the transport of conservation quantities can be described by the token game. Examples are the mass (in non-discrete form like water or in well defined units like goods), energy or information. Places should be used for system components which are able to store or to transform the tokens. For people who are concerned with modelling and simulation, it can be said that places are the parts of the system where it is necessary to specify differential equations for the temporal changing of the token quantity. Examples are vessels and tanks for mass tokens; heat exchangers or evaporators in the case of energy tokens; and waiting queues in information nets. Transitions are, in the ideal case, components like pipes or wires which only transport the tokens without changing their amount. In addition the flux relation (iii) of the net definition (Section 2) must be considered during the modelling process.

Therefore a lot of knowledge in the area of process engineering and some experience for the

modelling phase is needed. However the expense is advantageous: Equation (9) offers the detection of faults with very low time constants for all operational conditions of the systems. Moreover, a prognosis of the system's future behavior will be possible if a fault is detected, because of its property of being an initial value problem. Such features are not provided by other fault detection methods.

The pt net description allows a different perspective to initial and boundary problems: If equation (7) is true, a covering S -invariant will exist. Remember, this is a S -invariant which does not contain any other elements than 1. From this it follows that a system will be describable as an initial value problem if a covering S -invariant exists. If no such invariant can be evaluated the problem under consideration is a boundary value problem.

The example given in Section 4 is a typical case. In many industrial plants cooling loops are needed. A two phase cooling loop shows a possible extension for the Petri net system description. As previously mentioned in our example, all steam containing components have been modelled as transitions which do not influence the fault criterion (9). The consequence of this simplification is some additional noise in the residual of Fig. 2. More suited to the problem could be the choice of a higher class of Petri nets, the coloured nets. A difference to pt nets is their property of token individuality. More than one type of token can exist and one token type can be changed into another. In the cooling loop example there could be 2 types, the water and the steam tokens. Their transformation takes place in the steam generators and the condensers. But at the moment it is not possible to state if the coloured net description will be appropriate in practice in the area of online fault detection. This is because there does not exist a rich theory of this class of Petri nets and further research is required.

6. CONCLUSIONS

The description of technical systems in terms of the Petri net theory enables the formulation of a new online fault detection criterion. If the S -invariant of the system, which is evaluated in its initial state, does not hold at a subsequent time point, alarm will be given. This criterion will be applicable if the net token flux across the envelope of the system is zero mean and no tokens are lost because of firing. Under these assumptions the S -invariant describes the token conservation; therefore the method presented is applicable only to physical conservation quan-

tities. As a consequence of the assumptions named above there exists a covering S -invariant or, in other words, the system is an initial value problem. Thus the detection of abnormal process behavior or measurement faults with very low time constants becomes possible and a prognosis of the future system behavior can be given in the error case. But due to the simplicity of the fault detection criterion no diagnosis of the failure localization can be provided. The method presented is predestinated for the surveillance of complex technical systems like production lines or transport circuits, as it is demonstrated by an example. Because of the lack of the diagnosis feature this method should be considered as the online part of a process information system which is able to trigger a (possible offline) diagnosis and interpretation unit.

Acknowledgements—The author wishes to thank J. Brummer, GRS, for many helpful discussions.

REFERENCES

- Abel, D. (1987) Modellbildung und Analyse ereignisorientierter Systeme mit Petri-Netzen. *Fortschrittsberichte VDI*, Reihe 8, Nr. 142, VDI, Düsseldorf (in German).
- Basseville, M. and A. Benveniste (1986) *Detection of Abrupt Changes in Signals and Dynamical Systems*. Springer, Berlin.
- Pascoletti, K.-H. (1986) *Diophantische Systeme und Lösungsmethoden zur Bestimmung aller Invarianten in Petri-Netzen*. Oldenbourg, München (in German).
- Prock, J. (1988) Mathematical modeling of a steam generator for sensor fault detection. *Appl. Math. Modelling*, 12, 581–592.
- Prock, J. (1989a) Sensor fault detection in dynamical systems. *Proc. Early Failure Detection and Diagnosis in Nuclear Power Plants* (IAEA specialists' meeting) Dresden 265–274.
- Prock, J. (1989b) Ein prozessunabhängiges Konzept zur Meschleierkennung mit Hilfe analytischer Redundanzmethoden. Techn. report, GRS A 1620, Gesellschaft für Reaktorsicherheit, Garching (in German).
- Reisig, R. (1982) *Petri Nets*. Springer, Berlin.

An Extended Direct Scheme for Robust Adaptive Nonlinear Control*

I. KANELAKOPOULOS,†§ P. V. KOKOTOVIC† and R. MARINO†

A new direct adaptive control scheme is developed for nonlinear systems satisfying an extended matching condition, and shown to be robust with respect to unmodeled dynamics.

Key Words—Adaptive control, nonlinear systems, extended matching, robustness, unmodeled dynamics

Abstract—The proposed adaptive scheme achieves regulation for a class of nonlinear systems with unknown constant parameters and unmodeled dynamics. The scheme does not employ overparametrization and does not restrict the class of nonlinearities by any growth conditions. Instead, the dependence on the unknown parameters is restricted by an extended matching condition, which, however, is satisfied in many systems of practical importance, such as most types of electric motors.

1. INTRODUCTION

THE DIRECT adaptive regulation scheme of Taylor *et al.* (1989) is, whenever applicable, a simpler alternative to more elaborate schemes (Nani and Araposthathis, 1988; Pomet and Praly, 1989a; Sastry and Isidori, 1989). In addition to its simplicity, the direct scheme has also a robustness property with respect to unmodeled dynamics which are present in most engineering applications. An overview by Sastry and Kokotovic (1988) shows that similar robustness properties are yet to be established for other nonlinear adaptive schemes. Another major advantage of the simple direct scheme is its applicability to systems with nonlinearities which are not globally Lipschitz, like x^2 or x_1x_2 . In this regard, other schemes are more restrictive because they assume that the nonlinearities are globally Lipschitz or satisfy some "linear

growth" condition. To stress the practical importance of nonlinearities which are not globally Lipschitz, let us remind the reader that they are common in mechanical systems with centrifugal forces, in electrical systems with flux-current or flux-speed products, in chemical kinetics, etc. The most recent works of Pomet and Praly (1989b), Bastin and Campion (1989) and Campion and Bastin (1990) have made significant progress toward removing the global Lipschitz conditions for systems without unmodeled dynamics.

Unfortunately, the applicability of the simple direct scheme, unlimited by the type of nonlinearity, is limited in its dependence on the unknown constant parameters. While, as in most other schemes, this dependence is assumed to be linear, a further restriction is that the unknown parameters appear only in system equations with control variables. Only a narrow class of nonlinear systems satisfies this *strict matching condition* exactly. To broaden this class, Taylor *et al.* (1989) use matched reduced-order models and treat the unmatched terms as unmodeled dynamics. For example, a matched reduced-order model of an electric motor with uncertain load is only its mechanical equation, while all the electrical phenomena are to be treated as unmodeled dynamics.

This paper introduces an *extended matching condition* which further broadens the applicability of the simple direct scheme without sacrificing any one of its advantages. The unknown constant parameters are now allowed to appear also in equations separated from the control variables by one integration. As an illustration, an electrical equation can now be added to the above-mentioned model of an electric motor. This makes the effects of unmodeled dynamics less significant.

* Received 30 March 1989, revised 2 February 1990, received in final form 15 June 1990. The original version of this paper was presented at the IFAC Symposium on Nonlinear Control System Design which was held in Capri, Italy during June, 1989. The published proceedings of this IFAC Meeting may be ordered from Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor C. C. Hang, under the direction of Editor P. C. Parks.

† Coordinated Science Laboratory, University of Illinois, 1101 W. Springfield Ave., Urbana, IL 61801, USA.

‡ Dipartimento di Ingegneria Elettronica, Seconda Università di Roma "Tor Vergata", Rome 00173, Italy.

§ Author to whom all correspondence should be addressed.

The *extended direct scheme* presented in this paper is a nonlinear version of a direct scheme proposed for linear systems of relative degree two; see, e.g. the recent book by Narendra and Annaswamy (1989). However, there is a major difference between the linear and nonlinear versions. In the linear case, the relative-degree-two scheme is a step toward a more general scheme. In the nonlinear case no general scheme is yet available which can handle as broad a class of nonlinearities as the extended direct scheme.

The organization and style of this paper stress simplicity and applicability, rather than theoretical novelty, of the proposed extended direct scheme. Although the same scheme can be developed for multi-input systems (Kanellakopoulos *et al.*, 1989, 1990), the presentation is focused on the single-input case. As a further simplification, the scheme is developed for systems in a special form in which the meaning of the extended matching condition is obvious. A more general class of nonlinear systems, transformable into this special form, is defined in the Appendix. The stability proofs in this paper make use of Lyapunov functions which, since an early paper by Parks (1966), have been a standard adaptive control tool. The tutorial character of the paper is enhanced by a position control example which illustrates analytical derivations and gives a hint of potential applications.

2. THE EXTENDED DIRECT SCHEME

The nonlinear plant with an unknown constant parameter vector $a = [a_1, \dots, a_p]^T$ is assumed to be in the form

$$\dot{z} = f_0(z) + \sum_{i=1}^p a_i f_i(z) + g_0(z)u, \quad (2.1)$$

where $f_0, f_1, \dots, f_p, g_0$ are smooth vector fields on B_z and $g_0(z)$ is bounded away from zero on B_z , a subset of \mathbb{R}^n .

Proposition 1 in the Appendix gives necessary and sufficient conditions for the existence of a change of coordinates $x = \phi(z)$ and a state feedback control $u = \alpha(z) + \beta(z)v$, where v is a new control variable, which transform (2.1) into the special form

$$\dot{x}_i = x_{i+1}, \quad 1 \leq i \leq n-2$$

$$\dot{x}_{n-1} = x_n + \sum_{i=1}^p a_i w_{1i}(x) = x_n + a^T w_1(x) \quad (2.2)$$

$$\dot{x}_n = v + \sum_{i=1}^p a_i w_{2i}(x) = v + a^T w_2(x),$$

where the expressions for $w_1(x)$ and $w_2(x)$ are given by (A.12) and (A.13) in the Appendix.

Note that $w_1(x)$ and $w_2(x)$ are smooth vector fields on $B_x = \phi(B_z)$.

The meaning of the *extended matching condition* (A.3), formulated in the Appendix, is clearly displayed in (2.2). The original system (2.1) has been transformed into a chain of $n-2$ integrators and two nonlinear equations with unknown parameters. The control variable v enters only the last equation and, hence, the *strict matching condition* of Taylor *et al.* (1989) is not satisfied unless the vector field $w_1(x)$ is identically zero. In the case of extended matching $w_1(x)$ is not required to vanish and, hence, unknown parameters are allowed to appear in the last two equations of (2.2). Note also that none of the functions appearing in (2.1) and (2.2) are required to be globally Lipschitz or to satisfy some other linear growth or sector conditions.

Instead of the true parameter values $a = [a_1, \dots, a_p]^T$, which are unknown, a controller will be designed using parameter estimates $\hat{a} = [\hat{a}_1, \dots, \hat{a}_p]^T$. The first step in this direction is to introduce a new state \hat{x}_n , instead of x_n , using the expression

$$\hat{x}_n = x_n + \sum_{i=1}^p \hat{a}_i w_{1i}(x) = x_n + \hat{a}^T w_1(x). \quad (2.3)$$

Before we employ (2.3), we need to know that the mapping $x_n \rightarrow \hat{x}_n$ is one-to-one, onto and continuous. Hence, we assume that there exist $B_a \subset \mathbb{R}^p$ and a constant $\delta > 0$ such that

$$1 + \sum_{i=1}^p a_i \left| \frac{\partial w_{1i}(x)}{\partial x_n} \right| \geq \delta, \quad \forall x \in B_x, \quad \forall a \in B_a. \quad (2.4)$$

Then, the last two equations of (2.2) are rewritten as

$$\dot{\hat{x}}_{n-1} = \hat{x}_n + (a - \hat{a})^T w_1(x) \quad (2.5)$$

$$\begin{aligned} \dot{\hat{x}}_n &= [v + a^T w_2(x)] \left(1 + \hat{a}^T \frac{\partial w_1(x)}{\partial x_n} \right) \\ &\quad + \hat{a}^T \left(\sum_{i=1}^{n-1} \frac{\partial w_{1i}(x)}{\partial x_i} x_{i+1} + \frac{\partial w_1(x)}{\partial x_{n-1}} a^T w_1(x) \right) \\ &\quad + \hat{a}^T w_1(x) \\ &= [v + a^T w_2(x)] \beta(x, \hat{a}) + \hat{a}^T w_3(x) \\ &\quad + \hat{a}^T w_4(x) a + \hat{a}^T w_1(x). \end{aligned} \quad (2.6)$$

Note that $\beta(x, \hat{a})$ and the elements of $w_3(x)$, $w_4(x)$ defined by (2.6) are smooth functions on $B_x \times B_a$.

We now proceed to find a control which renders equations (2.5)–(2.6) linear in the parameter error $a - \hat{a}$ and makes them otherwise independent of the unknown parameter vector a . On closer inspection of (2.6), we see that these tasks will be accomplished by a control of the

form

$$v = -\hat{a}^T w_2(x) - \frac{1}{\beta(x, \hat{a})} [k_1 x_1 + \dots + k_{n-1} x_{n-1} + k_n (x_n + \hat{a}^T w_1(x)) + \hat{a}^T w_3(x) + \hat{a}^T w_4(x) \hat{a} + \hat{a}^T w_1(x)], \quad (2.7)$$

which, in addition to nonlinearity cancellation terms, contains a linear state feedback part with constant gains k_1, \dots, k_n . These gains are chosen to place at some desired stable locations the roots of the characteristic polynomial

$$s^n + k_n s^{n-1} + \dots + k_2 s + k_1 = 0. \quad (2.8)$$

To verify that the control law (2.7) is implementable on $B_x \times B_a$, first observe that $|\beta(x, \hat{a})| \geq \delta$ by assumption (2.4). Second, and this is a characteristic of the extended direct scheme, note the presence of the time derivative $\dot{\hat{a}}$ of the parameter estimate \hat{a} . It may appear that the implementation of (2.7) would require differentiators. Fortunately, this is not so. As we shall see, the parameter update law for \hat{a} will furnish $\dot{\hat{a}}$ as a known explicit function of available signals. To design this update law, we substitute the control (2.7) into (2.6) and obtain

$$\dot{\hat{x}}_n = [\beta(x, \hat{a}) w_2^T(x) + \hat{a}^T w_4(x)](a - \hat{a}) - k_1 x_1 - \dots - k_n \hat{x}_n \quad (2.9)$$

Then we introduce the following compact notation:

$$\hat{x} = \psi(x, \hat{a}) = \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_n + \hat{a}^T w_1(x) \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ \hat{x}_n \end{bmatrix}, \quad (2.10)$$

$$A = \begin{bmatrix} 0 & & & \\ \vdots & & & \\ 0 & & I & \\ -k_1 & \dots & -k_n \end{bmatrix}, \quad (2.10)$$

$$\begin{aligned} \tilde{W}(\hat{x}, \hat{a}) &= \tilde{W}(\psi(x, \hat{a}), \hat{a}) = W(x, \hat{a}) \\ &= \begin{bmatrix} 0 & \dots & 0 \\ 0 & \dots & 0 \\ w_1^T(x) \\ \beta(x, \hat{a}) w_2^T(x) + \hat{a}^T w_4(x) \end{bmatrix}. \end{aligned} \quad (2.11)$$

In this notation, the system with feedback control (2.7) is rewritten as

$$\dot{\hat{x}} = A\hat{x} + \tilde{W}(\hat{x}, \hat{a})(a - \hat{a}). \quad (2.12)$$

This "error form" shows that asymptotic stability is achieved when the estimate \hat{a} is correct, $a - \hat{a} = 0$, because (2.8) is the characteristic polynomial of A . We now prove that, even when $\hat{a} \neq a$, stability will be achieved with the

parameter update law

$$\dot{\hat{a}} = \Gamma \tilde{W}^T(\hat{x}, \hat{a}) P \hat{x} = \Gamma W^T(x, \hat{a}) P \psi(x, \hat{a}), \quad (2.13)$$

where Γ is a positive definite matrix ("adaptation gain") and $P > 0$ is chosen to satisfy the Lyapunov equation

$$PA + A^T P = -I. \quad (2.14)$$

To confirm that the update law (2.13) leads to a differentiator-free implementation of the feedback control (2.7), let us rewrite (2.7) as an explicit function of the available signals x and \hat{a} :

$$\begin{aligned} v &= -\hat{a}^T w_2(x) - \frac{1}{\beta(x, \hat{a})} \\ &\times [k_1 x_1 + \dots + k_n (x_n + \hat{a}^T w_1(x)) \\ &+ \hat{a}^T w_3(x) + \hat{a}^T w_4(x) \hat{a} \\ &+ w_1^T(x) \Gamma W^T(x, \hat{a}) P \psi(x, \hat{a})]. \end{aligned} \quad (2.15)$$

Our next task is to prove stability and a convergence property of the adaptive scheme consisting of the state equation (2.12) and the parameter update law (2.13).

Lemma 1. The equilibrium $\hat{x} = 0$, $\hat{a} = a$, of the scheme (2.12)–(2.13) is stable for every $a \in B_a$. Moreover, there exists a set $\hat{\Omega} \subset \mathbb{R}^{n+p}$ such that from all $(\hat{x}(0), \hat{a}(0)) \in \hat{\Omega}$ the state $\hat{x}(t)$ converges to zero.

$$\lim_{t \rightarrow \infty} \hat{x}(t) = 0. \quad (2.16)$$

Proof. Differentiating the Lyapunov function

$$V(\hat{x}, \hat{a}) = \hat{x}^T P \hat{x} + (a - \hat{a})^T \Gamma^{-1} (a - \hat{a}) \quad (2.17)$$

along the solutions of (2.12)–(2.13), we obtain

$$\begin{aligned} \dot{V} &= \hat{x}^T (A^T P + PA) \hat{x} + 2\hat{x}^T P \tilde{W}(\hat{x}, \hat{a})(a - \hat{a}) \\ &\quad - 2\dot{\hat{a}}^T \Gamma^{-1} (a - \hat{a}) \\ &= -\|\hat{x}\|^2 + 2[\hat{x}^T P \tilde{W}(\hat{x}, \hat{a}) \Gamma - \dot{\hat{a}}^T] \Gamma^{-1} (a - \hat{a}) \\ &= -\|\hat{x}\|^2 \leq 0. \end{aligned} \quad (2.18)$$

This proves the stability of the equilibrium $\hat{x} = 0$, $\hat{a} = a$. The convergence result (2.16) now follows from LaSalle's invariance theorem, by which $(\hat{x}(t), \hat{a}(t)) \rightarrow M$ as $t \rightarrow \infty$, where M is the largest invariant set of (2.12)–(2.13) contained in the set $\{(\hat{x}, \hat{a}) : \dot{V} = 0\}$ where $V = 0$. Finally, in view of (2.18), the function $V(\hat{x}(t), \hat{a}(t))$ is nonincreasing and, hence, a subset of $\hat{\Omega}$ is the set

$$\hat{\Omega}_c = \{(\hat{x}, \hat{a}) : V(\hat{x}, \hat{a}) \leq c\}, \quad (2.19)$$

where c is the largest constant such that

$$\Omega_c = \{(x, \hat{a}) : V(\psi(x, \hat{a}), \hat{a}) \leq c\} \subset B_x \times B_a. \quad (2.20)$$

□

Our final task is to prove the same stability and convergence properties for the actual adaptive system

$$\begin{aligned} \dot{x}_i &= x_{i+1}, \quad 1 \leq i \leq n-2 \\ \dot{x}_{n-1} &= x_n + a^T w_1(x) \\ \dot{x}_n &= (a - \hat{a})^T w_2(x) - \frac{1}{\beta(x, \hat{a})} \\ &\quad \times [k_1 x_1 + \dots + k_n x_n + k_n \hat{a}^T w_1(x) \\ &\quad + \hat{a}^T w_3(x) + \hat{a}^T w_4(x) \hat{a} \\ &\quad + w_1^T(x) \Gamma W^T(x, \hat{a}) P \psi(x, \hat{a})] \\ \dot{\hat{a}} &= \Gamma W^T(x, \hat{a}) P \psi(x, \hat{a}). \end{aligned} \quad (2.21)$$

This system differs from the scheme (2.12)–(2.13) only in the last state, which in the scheme is \hat{x}_n , while here it is x_n . To determine the equilibrium $x = x^*$, $\hat{a} = a$ of the system (2.21) which corresponds to the equilibrium $\hat{x} = 0$, $\hat{a} = a$ of the scheme (2.12)–(2.13), we note that at $x_1 = x_2 = \dots = x_{n-1} = 0$ the equation

$$\dot{x}_{n-1} = x_n + a^T w_1(x) = 0 \quad (2.22)$$

has, because of (2.4), a unique solution x_n^* for each $a \in B_a$. A direct substitution proves that

$$x = x^* = [0 \dots 0 \ x_n^*]^T, \quad \hat{a} = a \quad (2.23)$$

is the equilibrium of (2.21) corresponding to $\hat{x} = 0$, $\hat{a} = a$.

Theorem 1. The equilibrium $x = x^*$, $\hat{a} = a$ of the adaptive system (2.21) is stable for every $a \in B_a$ and the set Ω_V defined in (2.20) is a subset of its region of attraction. Moreover, for all $(x(0), \hat{a}(0)) \in \Omega_V$, the state $x(t)$ converges to its equilibrium value x^* , that is,

$$\lim_{t \rightarrow \infty} x(t) = x^*. \quad (2.24)$$

Proof. Lemma 1 proves the stability of $\hat{x} = 0$, $\hat{a} = a$. On the other hand, the mapping $(x, \hat{a}) \rightarrow (\hat{x}, \hat{a})$ is one-to-one, onto and continuous and it maps the point (x^*, a) to the point $(0, a)$. This proves the stability of $x = x^*$, $\hat{a} = a$, and, furthermore, implies that the solution $(x(t), \hat{a}(t))$ of (2.21) for any $(x(0), \hat{a}(0)) \in \Omega_V$ is uniformly bounded and remains in $B_x \times B_a$ for all $t \geq 0$. Differentiation of $\dot{x}_{n-1} = x_n + a^T w_1(x)$ in (2.2) and the use of (2.6) gives

$$\begin{aligned} \dot{x}_{n-1} &= [v + a^T w_2(x)] \beta(x, a) \\ &\quad + a^T w_3(x) + a^T w_4(x) a. \end{aligned} \quad (2.25)$$

Recall that all the functions appearing in (2.25) are smooth on $B_x \times B_a$, including, because of (2.4), the control v as expressed by (2.15). Therefore, $\dot{x}_{n-1}(t)$ is also uniformly bounded and $\dot{x}_{n-1}(t)$ is uniformly continuous. From (2.16) we

have

$$\lim_{t \rightarrow \infty} x_{n-1}(t) = x_{n-1}(0) + \lim_{t \rightarrow \infty} \int_0^t \dot{x}_{n-1}(\tau) d\tau = 0. \quad (2.26)$$

By Barbalat's lemma [see, e.g. Sastry and Bodson (1989, p. 19)], the uniform continuity and integrability of $\dot{x}_{n-1}(t)$ imply

$$\lim_{t \rightarrow \infty} \dot{x}_{n-1}(t) = \lim_{t \rightarrow \infty} (x_n(t) + a^T w_1(x(t))) = 0. \quad (2.27)$$

Combining (2.16), (2.22) and (2.27) with the fact that $w_1(x)$ is a smooth vector field, we conclude that

$$\lim_{t \rightarrow \infty} x_n(t) = x_n^*. \quad (2.28)$$

□

3. ROBUSTNESS TO UNMODELED DYNAMICS

The stability results in the preceding section assume that, up to a set of unknown constant parameters, an exact model of the plant is available. A crucial question to be addressed in this section is whether, and to what extent, these stability properties are preserved in the presence of unmodeled dynamics. As in some earlier works on adaptive control, such as Ioannou and Kokotovic (1983) and Taylor *et al.* (1989), the unmodeled dynamics are assumed to be fast and are treated as *singular perturbations*, that is, the change of model order is parameterized by a scalar μ . When $\mu > 0$, the unmodeled dynamics are included and the model order is higher than assumed in the adaptive scheme design. When $\mu = 0$, the unmodeled dynamics vanish and the model reduces to the assumed design model. In applications, the parameter μ is used to represent small inertias, inductances, time constants etc., as illustrated by many physical examples in Kokotovic *et al.* (1986).

Starting with the design model in the special form (2.2), we perturb it in the following way:

$$\begin{aligned} \dot{x}_i &= x_{i+1}, \quad 1 \leq i \leq n-2 \\ \dot{x}_{n-1} &= x_n + a^T w_1(x) \\ \dot{x}_n &= a^T w_2(x) + r^T(x) \xi + g_1(x) u \\ \mu \dot{\xi} &= Q(x) \xi + g_2(x) u, \end{aligned} \quad (3.1)$$

where $\xi \in \mathbb{R}^s$ is the state of the unmodeled dynamics, and the entries of $r(x)$, $Q(x)$, $g_1(x)$ and $g_2(x)$ are smooth functions on B_x . Within the framework of singular perturbations, the simple form of unmodeled dynamics included in (3.1) is sufficiently general, because, as shown by Kanellakopoulos (1989), other perturbed versions of the general model (2.1) can be transformed into (3.1). The relationship of (3.1)

with the unperturbed design model (2.2) will become clear after the change of variables

$$\xi = h(x, u) + \eta, \quad (3.2)$$

which exhibits the function

$$h(x, u) = -Q^{-1}(x)g_2(x)u \quad (3.3)$$

as the *quasi-steady-state* of ξ , and η as its *fast transient*. The unmodeled dynamics are assumed to be asymptotically stable for all fixed $x \in B_r$, that is, there exists $\sigma_1 > 0$ such that

$$\operatorname{Re} \lambda\{Q(x)\} \leq -\sigma_1, \quad \forall x \in B_r. \quad (3.4)$$

This, in turn, assures the existence of $Q^{-1}(x)$ in (3.3). Along with the change of state variables (3.2), we introduce the following change of the control variable

$$v = [g_1(x) - r^T(x)Q^{-1}(x)g_2(x)]u = g(x)u \quad (3.5)$$

and assume that there exists $\sigma_2 > 0$ such that

$$|g(x)| \geq \sigma_2 \quad \forall x \in B_r. \quad (3.6)$$

These changes of variables transform the perturbed model (3.1) into

$$\begin{aligned} \dot{x}_i &= x_{i+1}, \quad 1 \leq i \leq n-2 \\ \dot{x}_{n-1} &= x_n + a^T w_1(x) \\ \dot{x}_n &= v + a^T w_2(x) + r^T(x)\eta \\ \mu \dot{\eta} &= Q(x)\eta - \mu \hat{h}(x, \hat{a}, \eta, a), \end{aligned} \quad (3.7)$$

where the time derivative

$$\hat{h} = \left[h_1 - h_n \frac{v}{g}, g_1 \right] \hat{x} + \frac{1}{g} h_n v \quad (3.8)$$

is expressed as a function of x, \hat{a}, η and a . The expression for \hat{x} as a function of x, a, η and v is given by (3.7), while the expressions for v and \hat{v} as functions of x, \hat{a}, η and a can be obtained from (2.15) and (2.13).

It is now obvious that for $\eta = 0$, the perturbed model (3.7) reduces to the assumed design model (2.2). However, when $\mu > 0$, we cannot expect that $\eta(t) \equiv 0$ even if $\eta(0) = 0$, because, in general, $\hat{h}(x, \hat{a}, \eta, a)$ is not zero. As we shall see, the stability properties established by Theorem 1 will be preserved if the term $\mu \hat{h}(x, \hat{a}, \eta, a)$ is sufficiently small. The analysis leading to such a result retraces the derivations of Section 2 for the perturbed model (3.7).

The adaptive scheme for the perturbed model (3.7) employs the same control (2.15), the same update law (2.13) and the same notation (2.10), (2.11) as in Section 2. The resulting adaptive scheme with unmodeled dynamics is

$$\begin{aligned} \dot{\hat{x}} &= A\hat{x} + \hat{W}(\hat{x}, \hat{a})(a - \hat{a}) + \hat{R}(\hat{x}, \hat{a})\eta \\ \dot{\hat{a}} &= \Gamma \hat{W}^T(\hat{x}, \hat{a})P\hat{x} \\ \mu \dot{\eta} &= \hat{Q}(\hat{x}, \hat{a})\eta - \mu \hat{h}(\hat{x}, \hat{a}, \eta, a), \end{aligned} \quad (3.9)$$

where $\hat{Q}(\hat{x}, \hat{a}) = Q(x)$, $\hat{R}(\hat{x}, \hat{a}) = [0 \cdots 0 \ r(x)]^T$ and $\hat{h}(\hat{x}, \hat{a}, \eta, a) = h(x, \hat{a}, \eta, a)$. To verify that $\hat{x} = 0$, $\hat{a} = a$, $\eta = 0$ is an equilibrium of (3.9), we return to (3.3) and (3.8). Then, with the help of (2.15), we define

$$\hat{h}(\hat{x}, \hat{a}) = h\left(x, \frac{v}{g(x)}\right) \quad (3.10)$$

$$\hat{h}(\hat{x}, \hat{a}, \eta, a) = \hat{h}_1(\hat{x}, \hat{a})\hat{x} + \hat{h}_2(\hat{x}, \hat{a})\hat{a}$$

and see that $\hat{h} = 0$ at $\hat{x} = 0$, $\hat{a} = a$. On the other hand, for $\eta = 0$ the first two equations of (3.9) represent the unperturbed scheme (2.12)–(2.13), which has an equilibrium at $\hat{x} = 0$, $\hat{a} = a$. For simplicity, we now make the additional assumption that $w_1(0) = w_2(0) = 0$. Under this assumption $\hat{x} = 0$ implies $x = 0$, that is, the equilibria of the adaptive scheme (3.9) and of the adaptive system are the same, and, furthermore, $\hat{W}(0, \hat{a}) = 0$ for all $\hat{a} \in B_a$.

As in Taylor *et al.* (1989), the stability of the equilibrium $\hat{x} = 0$, $\hat{a} = a$, $\eta = 0$ of (3.9) will be investigated using the composite Lyapunov function

$$\begin{aligned} V_c(\hat{x}, \hat{a}, \eta) &= c_1[\hat{x}^T P \hat{x} + (a - \hat{a})^T \Gamma^{-1}(a - \hat{a})] \\ &\quad + c_2 \eta^T \hat{P}_f(\hat{x}, \hat{a})\eta, \end{aligned} \quad (3.11)$$

where c_1 and c_2 are positive constants and $\hat{P}_f(\hat{x}, \hat{a})$ is the positive definite solution of

$$\hat{P}_f(\hat{x}, \hat{a})\hat{Q}(\hat{x}, \hat{a}) + \hat{Q}^T(\hat{x}, \hat{a})\hat{P}_f(\hat{x}, \hat{a}) = -I. \quad (3.12)$$

The time derivative of V_c along the solutions of (3.9) is

$$\begin{aligned} \dot{V}_c &= c_1[-\hat{x}^T \hat{x} + 2\hat{x}^T P \hat{R}(\hat{x}, \hat{a})\eta] \\ &\quad + c_2 \frac{1}{\mu} \eta^T \eta + \eta^T \hat{P}_f(\hat{x}, \hat{a}, \eta, a)\eta \\ &\quad - 2\eta^T \hat{P}_f(\hat{x}, \hat{a})\hat{h}(\hat{x}, \hat{a}, \eta, a) \end{aligned} \quad (3.13)$$

The function \hat{h} , as defined in (3.10), satisfies $\hat{h}(0, \hat{a}, 0, a) = 0$ for all $a, \hat{a} \in B_a$. Hence, it is bounded by

$$\|\hat{h}(\hat{x}, \hat{a}, \eta, a)\| \leq \rho_1 \|\hat{x}\| + \rho_2 \|\eta\| \quad (3.14)$$

where the constants ρ_1, ρ_2 are such that for all $x \in B_r$, $\hat{a} \in B_a$, $a \in B_a$ the following inequalities hold:

$$\|\hat{h}_1(\hat{x}, \hat{a})\hat{W}(\hat{x}, \hat{a})(a - \hat{a})\| \leq \rho_0 \|\hat{x}\| \quad (3.15)$$

$$\|\hat{h}_1(\hat{x}, \hat{a})A + \hat{h}_2(\hat{x}, \hat{a})\Gamma \hat{W}^T(\hat{x}, \hat{a})P\| + \rho_0 \leq \rho_1 \quad (3.16)$$

$$\|\hat{h}_1(\hat{x}, \hat{a})\hat{R}(\hat{x}, \hat{a})\| \leq \rho_2. \quad (3.17)$$

Furthermore, we choose constants c_1, c_2 and c_3

which satisfy

$$2 \|\hat{P}_f(\hat{x}, \hat{a})\| \rho_1 \leq c_1 \quad (3.18)$$

$$2 \|\hat{P}\hat{R}(\hat{x}, \hat{a})\| \leq c_2 \quad (3.19)$$

$$2 \|\hat{P}_f(\hat{x}, \hat{a})\| \rho_2 + \|\hat{P}_f(\hat{x}, \hat{a}, \eta, a)\| \leq c_3 \quad (3.20)$$

for all $x \in B_x$, $\hat{a} \in B_a$, $\eta \in B_\eta$, $a \in B_a$.

A stronger effect of the fast variable η manifests itself as an increase in the values of c_1 , c_2 and c_3 . This is particularly clear for c_2 , because c_2 bounds the matrix \hat{R} through which η enters into the adaptive scheme (3.9). Constants c_1 and c_3 show the effects of η through the properties of the Lyapunov matrix \hat{P}_f . Using the bounds (3.14)–(3.20), we obtain from (3.13):

$$\dot{V}_t \leq -\begin{bmatrix} \|\hat{x}\| & \|\eta\| \end{bmatrix} \begin{bmatrix} c_1 & -c_1 c_2 \\ -c_1 c_2 & c_2 \left(\frac{1}{\mu} - c_3 \right) \end{bmatrix} \begin{bmatrix} \|\hat{x}\| \\ \|\eta\| \end{bmatrix} \quad (3.21)$$

This bound leads to the following robustness result:

Theorem 2. The equilibrium $x = 0$, $\hat{a} = a$, $\eta = 0$ of the perturbed adaptive system (3.7) with the feedback (2.15) and the update law (2.13) is stable for every $a \in B_a$ and for every μ satisfying

$$0 < \mu < \mu^* = \frac{1}{c_1 c_2 + c_3} \quad (3.22)$$

with c_1 , c_2 , c_3 as defined by (3.18)–(3.20). A subset of its region of attraction is the set

$$\Omega_c = \{(x, \hat{a}, \eta) : V_t(\psi(x, \hat{a}), \hat{a}, \eta) \leq c\}, \quad (3.23)$$

where c is the largest constant such that $\Omega_c \subset B_x \times B_a \times B_\eta$. Moreover, for all $(x(0), \hat{a}(0), \eta(0)) \in \Omega_c$ and $\mu \in (0, \mu^*)$, the state $(x(t), \eta(t))$ converges to its equilibrium value, that is

$$\lim_{t \rightarrow \infty} x(t) = 0, \quad \lim_{t \rightarrow \infty} \eta(t) = 0. \quad (3.24)$$

Proof. If (3.22) is satisfied, the matrix in (3.21) is positive definite and thus \dot{V}_t is negative semidefinite. This proves the stability of the equilibrium $\hat{x} = 0$, $\hat{a} = a$, $\eta = 0$ of the perturbed adaptive scheme (3.9). Hence, the equilibrium $x = 0$, $\hat{a} = a$, $\eta = 0$ of the perturbed adaptive system is stable and an estimate of its region of attraction is the set Ω_c in (3.23). Furthermore, the largest invariant set of (3.9), contained in the set where $\dot{V}_t = 0$, is the set $\{(\hat{x}, \hat{a}, \eta) : \hat{x} = 0, \eta = 0\}$. The convergence result (3.24) then follows from LaSalle's invariance theorem and the fact that $\lim_{t \rightarrow \infty} \hat{x}(t) = 0$ implies $\lim_{t \rightarrow \infty} x(t) = 0$. \square

It may appear that Theorem 2 requires a detailed knowledge of the unmodeled dynamics. In fact, even when nothing else is known about the unmodeled dynamics other than that they are fast and stable in the sense of (3.4), Theorem 2 still provides conceptual robustness bounds. Whatever the unmodeled dynamics, some strictly positive constants c_1 , c_2 , c_3 and, hence, μ^* exist so that the stability of the closed-loop system and the regulation property (3.24) are preserved for all $\mu \leq \mu^*$.

In applications, a more detailed description of the unmodeled dynamics may be available. For example, the dynamics of some known parts of the system (e.g. actuators and sensors) may have to be neglected in order to make the system appear in the special form (2.2). (In the next section, a motor position control system is brought to the special form (2.2) by neglecting the power amplifier dynamics.)

For a specific application, the time constant μ is known and the bounds (3.14)–(3.22) can be used to first determine a region $B_x \times B_a \times B_\eta$, and then evaluate the set Ω_c in which the proposed adaptive design is applicable. On the other hand, if Ω_c is a design specification, then the above bounds can be used to determine the required time constant μ . It is pointed out, however, that these bounds are conservative. Following the outlined procedure, tighter bounds can be obtained by taking into account details of the problem at hand.

4. DISCUSSION AND EXAMPLES

In this section we stress the practicality of the results in the preceding two sections by applying them to a common control problem. We then discuss a further extension of these results.

Example 1. *Position control of a DC-motor with uncertain load torque and unmodeled dynamics.* In normalized units, the position control system is described by

$$\begin{aligned} \frac{d\theta}{dt} &= \omega \\ \frac{d\omega}{dt} &= i + \sum_{j=1}^p a_j w_{1j}(\theta, \omega) \\ \frac{di}{dt} &= \frac{T_m}{T_e} (-\omega - i + \xi) \\ \mu \frac{d\xi}{dt} &= -\xi + g u, \end{aligned} \quad (4.1)$$

where θ , ω , i and ξ are the motor position, speed, and armature current and voltage, respectively, while the control variable is the input u into the amplifier with constant gain g .

The uncertain load torque is represented by a weighted sum of some known nonlinear functions $w_{ij}(\theta, \omega)$ with unknown constant weights a_j , $j = 1, \dots, p$. For illustrative purposes we take $p = 1$ and $w_{11}(\theta, \omega) = \omega^2$. The ratio T_e/T_m of the motor electrical and mechanical time constants was neglected in Taylor *et al.* (1989) in order to satisfy the strict matching condition. Here this quantity is retained in the model, and the singular perturbation parameter is the amplifier time constant μ .

The control objective is to regulate the motor position and speed to the desired constant values $\theta = \theta_{des}$ and $\omega = 0$. Denoting $x_1 = \theta - \theta_{des}$, $x_2 = \omega$, $x_3 = i$, we rewrite (4.1) in the form of the perturbed model (3.1):

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 + a_1 x_2^2 \\ \dot{x}_3 &= \frac{T_m}{T_e}(-x_2 - x_3 + \xi) \\ \mu \dot{\xi} &= -\xi + gu.\end{aligned}\quad (4.2)$$

The change of variables (3.2) is now $\xi = gu + \eta$, so that $h = gu$, $\dot{h} = g\dot{u}$, and the perturbed model (4.2), rewritten in the form (3.7), becomes

$$\begin{aligned}\dot{\hat{x}}_1 &= \hat{x}_2 \\ \dot{\hat{x}}_2 &= \hat{x}_3 + a_1 \hat{x}_2^2 \\ \dot{\hat{x}}_3 &= \frac{T_m}{T_e}(-\hat{x}_2 - \hat{x}_3 + gu + \eta) = v + \frac{T_m}{T_e} \eta \\ \mu \dot{\eta} &= -\eta - \mu g \dot{u}.\end{aligned}\quad (4.3)$$

As our design model (2.2) we use the first three equations with $\eta = 0$. The meaning of this design model is that the amplifier is replaced by its constant gain g and the voltage gu is applied to the motor armature. The design now proceeds as in Section 2. Using

$$\hat{\hat{x}}_3 = \hat{x}_3 + \hat{a}_1 \hat{x}_2^2 \quad (4.4)$$

and noting that assumption (2.4) is trivially satisfied, we design the control (2.7):

$$v = -[k_1 \hat{x}_1 + k_2 \hat{x}_2 + k_3 \hat{\hat{x}}_3 + 2\hat{a}_1 \hat{x}_2 \hat{\hat{x}}_3 + \hat{a}_1 \hat{x}_2^3]. \quad (4.5)$$

This results in the update law (2.13) with

$$\begin{aligned}\hat{W}(\hat{x}, \hat{a}_1) &= W(x, \hat{a}_1) = \\ &= \begin{bmatrix} 0 \\ 2\hat{a}_1 \hat{x}_2^3 \end{bmatrix}\end{aligned}\quad (4.6)$$

The perturbed adaptive scheme (3.9) is

$$\begin{aligned}\dot{\hat{x}}_1 &= \hat{x}_2 \\ \dot{\hat{x}}_2 &= \hat{x}_3 + (a_1 - \hat{a}_1) \hat{x}_2^2 \\ \dot{\hat{x}}_3 &= -k_1 \hat{x}_1 - k_2 \hat{x}_2 - k_3 \hat{\hat{x}}_3 + (a_1 - \hat{a}_1) \\ &\quad \times 2\hat{a}_1 \hat{x}_2^3 + \frac{T_m}{T_e} \eta \\ \dot{\hat{a}}_1 &= \Gamma \hat{W}^T(\hat{x}, \hat{a}_1) P \hat{x} \\ \mu \dot{\eta} &= -\eta - k \dot{u}.\end{aligned}\quad (4.7)$$

The largest invariant set of the scheme (4.7), contained in the set where $\dot{V}_1 = 0$ (see (3.21)), is the set $\{(\hat{x}, \hat{a}_1, \eta) : \hat{x} = 0, \eta = 0\}$. Thus, our scheme achieves regulation of the state $(\hat{x}(t), \eta(t))$.

Remark 1. The adaptive scheme presented in this paper can be further extended to include the case where unknown parameters enter the control vector field. (In the above example, this would occur if the ratio $T_e/T_m = a_1$ were also unknown.) A detailed analysis of this was presented in Kanellakopoulos *et al.* (1989). The main difference from the case treated in this paper is that the update law (2.13) depends on the control variable v . Therefore, v and \hat{a} are defined *implicitly*, and can be expressed as explicit functions of \hat{x} and \hat{a} only in a region S around the equilibrium $\hat{x} = 0$, $\hat{a} = a$. The boundaries of this solvability region are the manifolds on which the Jacobian of the implicit function defining v is singular. As an illustration, consider the following example.

Example 2. The system

$$\begin{aligned}\dot{\hat{x}}_1 &= (10 + a) \hat{x}_2 \\ \dot{\hat{x}}_2 &= -\hat{x}_2^3 + (10 + a) v\end{aligned}\quad (4.8)$$

is controllable for $a \neq -10$. Following the development of Section 2, we use

$$\hat{\hat{x}}_2 = (10 + \hat{a}) \hat{x}_2 \quad (4.9)$$

$$\begin{aligned}v &= \frac{1}{(10 + \hat{a})^2} [-\hat{x}_1 - (10 + \hat{a}) \hat{x}_2 \\ &\quad - \hat{a} \hat{x}_2 + (10 + \hat{a}) \hat{x}_2^3]\end{aligned}\quad (4.10)$$

and reduce (4.8) to the error form (2.12)

$$\begin{aligned}\dot{\hat{x}}_1 &= \hat{x}_2 + x_2(a - \hat{a}) \\ \dot{\hat{x}}_2 &= -\hat{x}_2 - \hat{x}_2^2 + (10 + \hat{a})v(a - \hat{a}),\end{aligned}\quad (4.11)$$

which is exponentially stable for $\hat{a} = a$. The Lyapunov function

$$V(\hat{x}, \hat{a}) = 1.5 \hat{x}_1^2 + \hat{x}_1 \hat{x}_2 + \hat{x}_2^2 + (a - \hat{a})^2 \quad (4.12)$$

results in the update law

$$\begin{aligned}\dot{\hat{a}} &= 1.5 \hat{x}_1 \hat{x}_2 + 0.5(10 + \hat{a}) \hat{x}_2^2 + (10 + \hat{a}) \\ &\quad \times [0.5 \hat{x}_1 + (10 + \hat{a}) \hat{x}_2] v.\end{aligned}\quad (4.13)$$

Equations (4.10) and (4.13) implicitly define v and \hat{a} as functions of x and \hat{a} . Eliminating \hat{a} , we get

$$(10 + \hat{a})[(10 + \hat{a}) + x_1(0.5x_1 + (10 + \hat{a})x_2)]v \\ = -x_1 - (10 + \hat{a})x_2 - 1.5x_1x_2^2 \\ + 0.5(10 + \hat{a})x_2^3. \quad (4.14)$$

To obtain v explicitly, the term multiplying it must be nonzero, which means that the adaptive scheme is defined only in a solvability region $S \subset B_1 \times B_u$. By inspection of (4.14), a subset of S is, for example

$$((x_1, x_2, \hat{a}) : |x_1| < 2, |5 + \hat{a}| < 3). \quad (4.15)$$

□

Remark 1 (cont'd). In general, the solvability condition for systems mentioned in Remark 1 requires that in the formulation of all the results of this paper, starting with the region of attraction estimate Ω_1 in (2.20), the set $B_1 \times B_u$ be replaced by the solvability region $S \subset B_1 \times B_u$. An approach which avoids the implicit definition of v and \hat{a} was proposed by Pomet and Praly (1989b). For (4.8) this approach consists in treating the parameter a appearing in the term $(10 + a)v$ as a second unknown parameter b to be estimated separately as \hat{b} . As a result of this overparametrization, the update law for \hat{a} no longer depends on v , which now appears in the definition of \hat{b} only.

Remark 2. For simplicity, all the results in this paper are given for the regulation problem. However, with standard technical modifications, the extended direct scheme can be used for the tracking of a given reference $y_{\text{des}}(t)$ by the output $y = x_1$ of the system (2.2). Moreover, zero dynamics can be added to (2.2) under the same assumptions as in Sastry and Isidori (1989). Of course, the invariance theorem of LaSalle does not apply to this time-varying problem, and the proofs of the tracking analogs of Lemma 1 and Theorem 1 employ the standard uniform continuity argument and Barbalat's lemma. Details and further extensions can be found in Kanellakopoulos (1989). To achieve robustness of the tracking scheme to unmodeled dynamics and bounded disturbances, one of the update law modifications proposed in adaptive linear control, such as the σ -modification of Ioannou and Kokotovic (1983), can be used. If this simple σ -modification scheme is used, robustness properties analogous to those established for the linear case in Theorems 5.2.1 and 6.5.1 of Ioannou and Kokotovic (1983) can be shown to hold in the nonlinear case.

5. CONCLUSIONS

The goals of the adaptive scheme proposed in this paper—simplicity, applicability to a wider class of nonlinearities and robustness to unmodeled dynamics—have been achieved under the extended matching condition. Although this condition restricts the dependence on the unknown parameters, it is satisfied by many systems appearing in engineering applications. In a manner similar to that illustrated by Example 1, the extended adaptive scheme is applicable to switched reluctance (Taylor, 1988) and permanent magnet stepper (Bodson and Chiasson, 1989) motors, and to other electromechanical or electrohydraulic systems with load and power input uncertainties. It would seem, therefore, that when the extended direct scheme is applicable, it should be given preference over more elaborate schemes. For problems to which this scheme is not applicable, the choice, or the development, of an appropriate scheme is a topic of ongoing research (Sastry and Isidori, 1989; Pomet and Praly, 1989a,b; Bastin and Campion, 1989; Campion and Bastin, 1990).

Acknowledgements.—This work was supported in part by the National Science Foundation under Grant ECS 88-18166 and in part by the Air Force Office of Scientific Research under Grant AFOSR 90-0011.

REFERENCES

- Bodson, M. and J. Chiasson (1989). Application of nonlinear control methods to the positioning of a permanent magnet stepper motor. *Proc. 28th CDC*, Tampa, FL, 531–532.
- Bastin, G. and G. Campion (1989). Indirect adaptive control of linearly parametrized nonlinear systems. *Preprints of the 3rd IFAC Symposium on Adaptive Systems in Control and Signal Processing*, Glasgow, UK.
- Campion, G. and G. Bastin (1990). Indirect adaptive state feedback control of linearly parametrized nonlinear systems. *Int. J. Adaptive Control Signal Proc.*, **4**, 345–358.
- Ioannou, P. A. and P. V. Kokotovic (1983). *Adaptive Systems with Reduced Models*. Springer, New York.
- Kanellakopoulos, I. (1989). Adaptive feedback linearization, stability and robustness. M.S. Thesis, University of Illinois, Urbana.
- Kanellakopoulos, I., P. V. Kokotovic and R. Marino (1989). Robustness of adaptive nonlinear control under an extended matching condition. *Preprints IFAC Symp. on Nonlinear Control System Design*, Capri, Italy, 192–197.
- Kanellakopoulos, I., P. V. Kokotovic and R. Marino (1990). Robust adaptive nonlinear control under extended matching conditions. Technical Report UIUC-ENG-90-2202 (DC-115), Coordinated Science Laboratory, University of Illinois, Urbana.
- Kokotovic, P. V., H. K. Khalil and J. O'Reilly (1986). *Singular Perturbation Methods in Control, Analysis and Design*. Academic Press, New York.
- Nam, K. and A. Arapostathis (1988). A model reference adaptive control scheme for pure-feedback nonlinear systems. *IEEE Trans. Aut. Control*, **AC-33**, 803–811.
- Narendra, K. S. and A. M. Annaswamy (1989). *Stable Adaptive Systems*. Prentice-Hall, Englewood Cliffs, NJ.
- Parks, P. C. (1966). Lyapunov redesign of model reference adaptive control systems. *IEEE Trans. Aut. Control*, **AC-11**, 362–367.
- Pomet, J. B. and L. Praly (1989a). Adaptive nonlinear control: an estimation-based algorithm. In Descusse, J.,

M. Fliess, A. Isidori and D. Leborgne (Eds), *New Trends in Nonlinear Control Theory*, Springer, Berlin.

Pomet, J. B. and L. Praly (1989b) Adaptive nonlinear regulation: equation error from the Lyapunov equation. *Proc. 28th CDC*, Tampa, FL, 1008-1013.

Sastry, S. S. and M. Bodson (1989) *Adaptive Control: Stability, Convergence and Robustness*, Prentice-Hall, Englewood Cliffs, NJ.

Sastry, S. S. and A. Isidori (1989) Adaptive control of linearizable systems. *IEEE Trans. Aut. Control*, **AC-34**, 1123-1131.

Sastry, S. S. and P. V. Kokotovic (1988) Feedback linearization in the presence of uncertainties. *Int. J. Adaptive Control Signal Proc.*, **2**, 327-346.

Su, R. (1982) On the linear equivalents of nonlinear systems. *Syst. Control Lett.*, **2**, 48-52.

Taylor, D. G. (1988) Feedback control of uncertain nonlinear systems with applications to electric machinery and robotic manipulators. Ph.D. Thesis, University of Illinois, Urbana.

Taylor, D. G., P. V. Kokotovic, R. Marino and I. Kanellakopoulos (1989) Adaptive regulation of nonlinear systems with unmodeled dynamics. *IEEE Trans. Aut. Control*, **34**, 405-412.

APPENDIX: THE EXTENDED MATCHING CONDITION

Consider the system

$$\dot{z} = f_0(z) + \sum_{i=1}^p a_i f_i(z) + u(t)g_0(z), \quad z \in \mathbb{R}^n \quad (A.1)$$

with $f_0(0) = 0$, $g_0(0) \neq 0$, $f_0, f_1, \dots, f_p, g_0$ smooth vector fields in U_1 , a neighborhood of the origin.

Feedback linearization condition The distributions

$$\mathcal{H}_i = \text{span}\{g_{0i}, ad_{f_0}g_{0i}, \dots, ad_{f_0}^{i-1}g_{0i}\}, \quad 0 \leq i \leq n-1 \quad (A.2)$$

are involutive and of constant rank $i+1$ in U_1 .

Extended matching condition

$$f_i \in \mathcal{H}_i, \quad 1 \leq i \leq p. \quad (A.3)$$

Proposition 1 There exists a state feedback control $u = \alpha(z) + \beta(z)v$, $\beta(z) \neq 0$ in U_1 , and a state space change of coordinates $x = \phi(z)$ in $U \subset U_1$, with $\phi(0) = 0$, such that the system

$$\dot{z} = f_0(z) + g_0(z)\alpha(z) + g_0(z)\beta(z)v + \sum_{i=1}^p a_i f_i(z) \quad (A.4)$$

becomes in the x coordinates

$$\begin{aligned} \dot{x}_i &= x_{i+1}, \quad 1 \leq i \leq n-2 \\ \dot{x}_{n-1} &= x_n + \sum_{i=1}^p a_i w_{1i}(x) \\ \dot{x}_n &= v + \sum_{i=1}^p a_i w_{2i}(x) \end{aligned} \quad (A.5)$$

if and only if the feedback linearization (A.2) and the extended matching (A.3) conditions are satisfied

Proof

Sufficiency It is proven in Su (1982) that condition (A.2) is sufficient for the existence of a function $\phi_1(z): U \rightarrow \mathbb{R}$ with the properties

- (1) $(d\phi_1, ad_{f_0}^{n-1}g_0) \neq 0$ in U , $(d\phi_1, X) = 0 \forall X \in \mathcal{H}_{n-2}$.
- (2) $(\phi_1, L_{g_0}\phi_1, \dots, L_{g_0}^{n-1}\phi_1)^T = (\phi_1, \dots, \phi_n)^T = \phi^T(z)$ is a change of coordinates in U , such that the system

$$\dot{z} = f_0(z) + u(t)g_0(z) \quad (A.6)$$

becomes in the new coordinates

$$\begin{aligned} \dot{x}_i &= x_{i+1}, \quad 1 \leq i \leq n-1 \\ \dot{x}_n &= L_{g_0}^n \phi_1(z) + L_{g_0}^{n-1} \phi_1(z)u(t) \end{aligned} \quad (A.7)$$

By using the state feedback (recall that $L_{g_0}L_{f_0}^{n-1}\phi_1(z) \neq 0$ in U)

$$u = \frac{1}{L_{g_0}L_{f_0}^{n-1}\phi_1(z)}(-L_{f_0}^n\phi_1(z) + v) = \alpha(z) + \beta(z)v, \quad (A.8)$$

the system (A.6) becomes

$$\dot{x}_i = x_{i+1}, \quad 1 \leq i \leq n-1 \quad (A.9)$$

In the new coordinates $x = \phi(z)$, the distribution \mathcal{H}_i is

$$\mathcal{H}_i = \text{span} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad (A.10)$$

and condition (A.3) implies that

$$f_i(x) = [0 \dots 0 \ w_{1i}(x) \ w_{2i}(x)]^T, \quad 1 \leq i \leq p, \quad (A.11)$$

where

$$w_{1i}(x) = L_{f_i}L_{g_0}^{i-1}\phi_1(\phi^{-1}(x)) \quad (A.12)$$

$$w_{2i}(x) = L_{f_i}L_{g_0}^{i-1}\phi_1(\phi^{-1}(x)) \quad (A.13)$$

In conclusion, the state feedback (A.8), applied to the system (A.1), results in

$$\begin{aligned} f_0(z) &= \frac{L_{f_0}^n \phi_1(z)}{L_{g_0}L_{f_0}^{n-1}\phi_1(z)}g_0(z) + L_{g_0}L_{f_0}^{n-1}\phi_1(z)g_0(z)v \\ &+ \sum_{i=1}^p a_i f_i(z) \end{aligned} \quad (A.14)$$

When expressed in the new coordinates $x = \phi(z)$, the system (A.14) becomes (A.5).

Necessity If there exist a state feedback and a change of coordinates transforming the system (A.1) into the system (A.5), one can directly verify that the conditions (A.2) and (A.3) are satisfied for the system (A.5). Since the conditions (A.2) and (A.3) are invariant under state feedback transformations and changes of coordinates, they are satisfied for the system (A.1). \square

A New Stable Compensator Design for Exact and Approximate Loop Transfer Recovery*

B. M. CHEN,[†] A. SABERI[‡] and P. SANNUTI[‡]

For minimum phase MIMO systems, a LTR design is presented via a new compensator which is stable and which allows much lower gain and thus lower controller band-width than the conventional observer based controller.

Key Words—Loop transfer recovery, robust control, linear quadratic Gaussian theory

1—In this paper, a new compensator structure for loop transfer recovery (LTR) is proposed. The proposed compensator (a) is open-loop stable, (b) guarantees closed-loop stability and above all (c) requires much smaller values of gain than the conventional observer based controller for the same degree of loop transfer recovery. The fact that the new compensator requires much smaller values of gain than the conventional controller results in several practical advantages, the most important among them being the reduction in controller band-width and freedom from the woes of saturation. The trade-off between the value of gain and the degree of loop transfer recovery as well as the bounds on singular values of sensitivity and complementary sensitivity functions is shown clearly. Both full and reduced order compensators for LTR when the design specifications are reflected either at the input or at the output point of the given plant are considered. Numerical examples illustrate the advantages of the new compensator structure.

The new compensator structure is inspired by a careful and clear understanding of how loop transfer recovery occurs when conventional observer based controllers are used. To motivate and deduce our new compensator structure, a unified treatment of observer theory for LTR is presented. In the context of such a unification, some new results are also given.

1. INTRODUCTION AND PROBLEM STATEMENT
IN MULTI-INPUT and multi-output (MIMO) feedback control system design, performance specifications such as command following, disturbance rejection, closed-loop band-width, stability robustness with respect to unstructured dynamic uncertainties etc., are naturally posed in the frequency domain in terms of sensitivity and complementary sensitivity functions. These sen-

sitivity and complementary sensitivity functions are related to the loop transfer matrices evaluated by breaking the control loop at critical points, commonly either the input or output point of the given plant. Recent results have shown that the formal mathematical synthesis procedures based on linear quadratic Gaussian (LQG) with loop transfer recovery (LTR), the so called LQG/LTR techniques, provide a broad flexibility in achieving the necessary loop transfer matrices. LQG/LTR technique is essentially a two-step approach and involves two separate designs of a linear quadratic regulator and either an observer or a Kalman filter based controller (Athans, 1986; Stein and Athans, 1987). The exact design procedure depends on the point, either the input or the output point of the plant, where the loop is broken to evaluate the open-loop transfer matrices. We will first concentrate our discussion on the case when the loop is broken at the input point of the plant. Dual discussion can be given for the case when the loop is broken at the output point. Thus in the two step procedure of LQG/LTR, the first step of design involves loop shaping by state feedback design to obtain an appropriate loop transfer function, called the target loop transfer function. Such a loop shaping is an engineering art and often involves the use of linear quadratic regulator (LQR) design in which the cost matrices are used as free design parameters to generate the target loop transfer function and thus the desired sensitivity and complementary sensitivity functions. However, when such a feedback design is implemented via an observer (or Kalman filter) based controller that uses only the output feedback, the obtained loop transfer function, in general, is not the same as the target loop transfer function, unless proper care is taken in designing the observers. This is when

* Received 28 September 1989; revised 24 April 1990; received in final form 14 May 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor V. Kucera under the direction of Editor H. Kwakernaak.

[†] Department of Electrical and Computer Engineering, Washington State University, Pullman, WA 99164-2752, U.S.A.

[‡] Department of Electrical and Computer Engineering, P.O. Box 909, Rutgers University, Piscataway, NJ 08855-0909, U.S.A. Author to whom all correspondence should be addressed.

Equation (1.9) can be given a physical interpretation. Considering the observer based controller as a device with its output as \hat{u} and inputs as u and y , it is easy to see that

$$\hat{x}(s) = (\Phi^{-1} + KC)^{-1}Bu(s) + (\Phi^{-1} + KC)^{-1}Ky(s),$$

and

$$\begin{aligned}\hat{u}(s) &= -F\hat{x}(s) \\ &= -M(s)u(s) - F(\Phi^{-1} + KC)^{-1}Ky(s).\end{aligned}\quad (1.11)$$

In view of the condition (1.9), equation (1.11) implies that the output \hat{u} of the observer based controller does not entail the feedback from the control signal u . However, the transfer function from u to \hat{x} is in general non-zero even when $M(s)$ is zero.

In practice, the condition $M(j\omega) = 0$ cannot always be satisfied exactly. The only recourse is then to make the size of $M(j\omega)$ in some sense small for all ω . Let the gain K be parameterized in terms of a scalar or a vector parameter σ and be denoted by $K(\sigma)$. Thus for ALTRI, one needs to obtain a $K(\sigma)$ such that,

$$M(s) = F(\Phi^{-1} + K(\sigma)C)^{-1}B \rightarrow 0 \quad \text{pointwise in } s \text{ as } \sigma \rightarrow \infty \quad (1.12)$$

The condition (1.12) involves the state feedback gain F . However, in order to have the state feedback and observer designs to be independent of one another, one needs to require that

$$(\Phi^{-1} + K(\sigma)C)^{-1}B \rightarrow 0 \quad \text{pointwise in } s \text{ as } \sigma \rightarrow \infty, \quad (1.13)$$

which is a sufficient condition for (1.12). Essentially there exists three methods of obtaining such a $K(\sigma)$. In their seminal work Doyle and Stein (1979) and others later on (Madiwale and Williams, 1985; Sogaard-Andersen, 1987b; Matson and Maybeck, 1987) explored Kalman filter formalism (or asymptotic LQG theory) in which additional fictitious process noise of intensity proportional to σ is injected into the system through the input into the plant and then the gain $K(\sigma)$ is calculated by solving the resulting filter Riccati equations. Sogaard-Andersen (1987a) proposed observer eigenstructure assignment techniques. More recently, Saberi and Sannuti (1988) proposed an asymptotic pole placement method which generalizes the earlier methods while simplifying the computational task in obtaining $K(\sigma)$. In all these ALTRI design methods, $\|K(\sigma)\| \rightarrow \infty$ as $\sigma \rightarrow \infty$ so that (1.13) can be satisfied asymptotically. Thus all these are high-gain schemes. This implies that any ALTRI design scheme should include a trade-off between the required

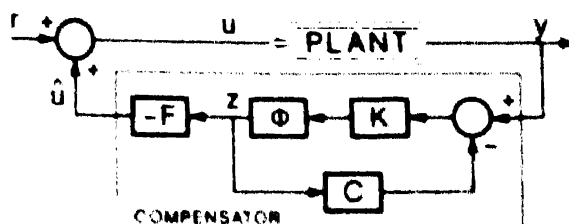


FIG. 3. Plant with full order compensator

robustness properties of the closed-loop system and the size of the feedback gain. The size of a feedback gain is very critical in many circumstances due to unavoidable controller bandwidth constraints. Thus a cursory exploration of the literature indicates a definite need to develop dynamic compensators which would preserve closed-loop stability and at the same time achieve ALTRI without requiring large amounts of gain. In this paper, we propose such a scheme of developing dynamic compensators of order either n (full order) or $n - m$ (reduced order).

Our central observation is this. When one is restricted to the framework of observer theory, the link from the control signal u to the observer via the control distribution matrix B is always present in the design configuration such as the one depicted in Fig. 2. In these observers, when $K(\sigma)$ is appropriately designed to achieve ALTRI, the effect of the above control-link on the output of observer based controller (namely \hat{u}) vanishes asymptotically as $\sigma \rightarrow \infty$. However, the effect of above link on \hat{x} in general is nonzero and hence the need for the above link in the conventional observers. Based on this discussion, we are inspired to remove the above mentioned link structurally right from the beginning of the design. In other words, to develop an appropriate compensator of order n , we consider the configuration illustrated in Fig. 3. Once the link from the control input to the controller or what is now called a compensator, is removed we embark on a new design philosophy which is outside the realm of observer theory and hence the separation principle is no longer valid. Without the backing or blessing of the separation principle, one has to prove that the design objectives of closed-loop stability and recovering the target loop shape can both be simultaneously achieved. We intend to do exactly this.

Our design philosophy is deceptively very simple. Except for structurally omitting the link mentioned earlier, our compensator is exactly the same as the conventional observer-based controller. We plan to obtain a $K(\sigma)$ such that (a) $A - K(\sigma)C$ has all its eigenvalues in the left half s plane (i.e. the compensator is open-loop stable) and that (b) the condition (1.12) is

satisfied asymptotically as $\sigma \rightarrow \infty$. For this purpose, we can use any of the existing methods of obtaining such a $K(\sigma)$. Thus our compensator design parallels in all respects the conventional observer design except for omitting the link mentioned earlier. Although our compensator structurally differs from the observer in a very simple way, it has a profound effect on the gain required for closed-loop stability and for ALTRI. We show theoretically that for the same gain, the difference between the target loop transfer function and the one achieved by our compensator is always much smaller than that that can be achieved by the observer based controller. But since our design method is also an asymptotic method, the above theoretical result does not reveal the whole story. The proof that our method works is evident from our examples. We have solved numerically many examples that appeared in the open literature, and noticed that the amount of gain required for the same degree of recovery by our compensator is orders of magnitude less than what is required by an observer-based controller. This obviously has a profound impact on the practical implementation of LQG/LTR schemes. Some specific attributes of our compensator are as follows:

1. Low values of gain obviously results in low compensator band-width, and hence much of the output noise that occurs at relatively high frequencies is filtered out. Furthermore, low values of gain relieves the design from ever present woes of saturation. To emphasize this, we refer to Sogaard-Andersen and Niemann (1989) who recently studied the design trade-offs between the level of loop transfer recovery and the necessary gain required by an observer-based controller. A major conclusion of their study is that the target loop transfer recovery design cannot always be achieved even when modest and practically meaningful constraints are imposed on the size of the observer gain. Furthermore, contrary to what has been discussed in the literature (e.g. Friedland, 1986; Baumgartner *et al.*, 1986), their study indicates that a high-gain from controller input to controller output affects the entire control-loop and in particular the control-noise signal ratio and the control-command signal ratio.

2. Since the given plant is of minimum phase, it is always possible to design an open-loop stable compensator to guarantee the over all closed-loop stability (Vidyasagar, 1985b). Our design results in an open-loop stable compensator. The advantages of having such a compensator cannot be over-emphasized. As is

known (Shaw, 1971), open-loop unstable compensators result in poor overall system sensitivity to plant parameter variations. Furthermore, physical realizability of open-loop unstable compensators is rather difficult.

Our discussion so far has been concerned with a full order observer-based controller and that too only with the case when the target open-loop transfer matrix is specified at the plant input point. Similar discussion pertains as well to other cases: (a) when a reduced order observer-based controller is considered and (b) when a target open-loop transfer matrix is specified at the plant output point. Even when a reduced order observer-based controller is used, we observe that exact or approximate LTR is possible if and only if the transfer function from the point where the input u of the plant is fed to the controller to the output point \hat{u} of the controller is either exactly or approximately zero. Thus again, our compensator structurally (i.e. physically) omits the link from the input point of the plant to the controller right from the beginning.

The paper is organized as follows. In Section 2, we consider conventional observers and while reviewing the existing theory, some clarifications and generalizations of it are presented. This section motivates the work that follows. Sections 3 and 4 respectively develop the full and reduced order compensators when the target open-loop transfer matrices are specified at the plant input point, while Section 5 dualizes the results of Sections 3 and 4 for the case when the target open-loop transfer matrices are specified at the plant output point. Section 6 deals with numerical results on some representative examples from the literature. Throughout this paper, A' denotes the transpose of A , I denotes an identity matrix while I_k denotes the identity matrix of dimension $k \times k$. $\lambda(A)$ and $\text{Re}[\lambda(A)]$ respectively denote the set of eigenvalues and real parts of eigenvalues of A . Similarly, $\sigma_{\max}[A]$ and $\sigma_{\min}[A]$ respectively denote the maximum and minimum singular values of A . The open left half plane is denoted by \mathcal{C}^- .

2. REVIEW OF LTR VIA OBSERVERS

The purpose of this section is to re-examine how LTR occurs when either full or reduced order observer-based controllers are used. This is done in order to obtain a better intuition and understanding of the theory of observer-based controllers for LTR so that our new compensator design, discussed in Sections 3 and 4, can easily be motivated and inspired. In this process of review, the conditions for achieving either exact or approximate LTR (ELTR or ALTR),

when either full or reduced order observer based controllers are used, are brought to the same frame work, i.e. the observer-based controller theory is unified into a single frame work. Such a review and unification leads to several new results. For example, no method of determining gain for full order observer-based controllers exists in the literature for the case of achieving ELTR. Here an explicit method of determining such a gain is given. Also, both the cases when the loop is broken at either the input or the output point of the plant are considered. In short, this section summarizes, generalizes, unifies and in some cases clarifies the existing results on LTR using conventional observer-based controllers.

2.1. Full order observers—ELTRI and ALTRI

We will first consider the full order observer-based controller when the target open-loop transfer matrix is evaluated when the loop is broken at the input point of the plant. $E_o(s)$, the error between the target loop transfer function $L(s)$ and that achievable by the observer based controller of Fig. 2 is given by (1.8). In the observer design, K is the only free design parameter. First of all in order to guarantee the closed-loop stability, K must be such that $A - KC$ is an asymptotically stable matrix, i.e.

$$\operatorname{Re}[\lambda(A - KC)] < 0. \quad (2.1)$$

The remaining freedom in choosing K can then be used to achieve LTR. In an attempt to find such a K , Doyle and Stein (1979) first gave a sufficient condition,

$$K(I + C\Phi K)^{-1}C\Phi B = B, \quad (2.2)$$

under which $E_o(j\omega) = 0$ for all ω . To understand the implications of (2.2), following Friedland (1986), we rewrite it in an equivalent way. In view of the identity,

$$\Phi K(I_p + C\Phi K)^{-1}C\Phi = \Phi - (\Phi^{-1} + KC)^{-1},$$

we have

$$\Phi K(I_p + C\Phi K)^{-1}C\Phi B = \Phi B - (\Phi^{-1} + KC)^{-1}B.$$

Thus, (2.2) implies that

$$(\Phi^{-1} + KC)^{-1}B = 0. \quad (2.3)$$

However, due to the nonsingularity of $(\Phi^{-1} + KC)^{-1}$, (2.3) implies that $B = 0$. But this is impossible in any real system. Thus the sufficient condition (2.2) cannot exactly be satisfied at all. However, (2.3) can asymptotically be satisfied for large gain without requiring $B = 0$.

The above discussion reveals that the condition (2.2) is poorly suited to study the loop transfer recovery problem. Realizing this,

Goodman (1984) proceeded to look at directly the error or mismatch function $E_o(s)$. Goodman studied square invertible systems. The following two lemmas represent minor extensions of Goodman's results and cover as well left invertible systems.

Lemma 1. $E_o(s)$, the error between the target loop transfer function $L(s)$ and that realized by the full order observer-based controller of Fig. 2 is given by

$$E_o(s) = M(s)(I_m + M(s))^{-1}(I_m + F\Phi B) \quad (2.4)$$

where

$$M(s) = F(\Phi^{-1} + KC)^{-1}B. \quad (2.5)$$

Lemma 2

$$E_o(j\omega) = 0 \quad \text{iff} \quad M(j\omega) = 0 \quad \text{for all} \quad \omega \in \Omega \quad (2.6)$$

where Ω is the set of all $0 < \omega < \infty$ for which $L_o(j\omega)$ and $L(j\omega)$ are well defined (i.e. all required inverses exist).

Thus equation (2.4) presents a clear perspective to study the basic mechanism by which both exact and approximate LTR occurs. It is clear that ELTRI is achievable if $M(j\omega) = 0$ exactly and on the other hand ALTRI is achievable if $\sigma_{\max}[M(j\omega)]$ can be made arbitrarily small for all ω . In order to investigate when $\sigma_{\max}[M(j\omega)]$ can be made either zero or arbitrarily small, assuming $A - KC$ is nondefective, Goodman [see also, Sogaard-Andersen, (1987c)] expands $M(s)$ in a dyadic form,

$$M(s) = \sum_{i=1}^n \frac{R_i}{s - \lambda_i} \quad (2.7)$$

where

$$R_i = FW_iV_i^HB$$

Here superscript H indicates the complex conjugate transpose. Also, W_i and V_i are respectively the right and left eigenvectors associated with an eigenvalue λ_i of $A - KC$ and they are scaled so that $WV^H = V^HW = I_n$ where

$$W = [W_1, W_2, \dots, W_n]$$

and

$$V = [V_1, V_2, \dots, V_n].$$

In view of Lemma 2, ELTRI is possible iff $M(j\omega) = 0$ for all ω . This is the case iff for each $i = 1$ to n , either $FW_i = 0$ or $V_i^HB = 0$ or both. Since F is designed to satisfy the required loop transfer function, $FW_i = 0$ is generically not satisfied. One can try to satisfy $V_i^HB = 0$ for as many indexes i as possible. However, it is not possible to design so that $V_i^HB = 0$ for all

indexes, $i = 1$ to n . Let us investigate how many left eigenvectors of $A - KC$ can satisfy $V_i''B = 0$. Let n_a and n_f be respectively the number of invariant zeros and infinite zeros (Sannuti and Saberi, 1987) of the given plant. In general $n_a \leq n - n_f$. We have the following result.

Lemma 3. For a left invertible plant, there exists a gain matrix K such that at most a total of $n - n_f$ left eigenvectors V_i , $i = 1$ to $n - n_f$, of $A - KC$ can satisfy the condition $V_i''B = 0$. Also, n_a of these $n - n_f$ eigenvalues are same as the invariant zeros of the plant while $n_f \approx n - n_f - n_a$ eigenvalues can be assigned freely. Furthermore, those left eigenvectors of $A - KC$ which are associated with the invariant zeros are same as the left zero directions of the plant.

Proof. See Shaked and Karcenas (1976); Saberi and Sannuti (1988).

Lemma 3 implies that the maximum number of indexes i that satisfy $V_i''B = 0$ is equal to the difference between the dynamic order and the number of infinite zeros of the given plant. The minimum number of infinite zeros is m and this happens for left invertible systems when all the infinite zeros are of order one, i.e. when CB is of maximum rank (Sannuti and Saberi, 1987). Then in view of Lemmas 2 and 3, ELTRI in general is impossible. It is possible only under some special circumstances. Goodman gives the following result.

Lemma 4. Let $A - KC$ be a nondefective matrix with left eigenvectors V_i , $i = 1$ to n , such that $V_i''B = 0$ for i equal to any $n - m$ distinct indexes among 1 to n . Then $E_0(j\omega) \approx 0$ for all $\omega \in \Omega$ is equivalent to $FB \approx 0$.

Lemmas 3 and 4 culminate in the following theorem.

Theorem 1. Consider the closed-loop system comprising of the plant and the full order observer-based controller as in Fig. 2. Then both asymptotic stability of the closed-loop system and ELTRI can be achieved under the following conditions:

1. $FB = 0$.
2. The given plant has all its infinite zeros of order one (i.e. CB is of maximal rank).
3. The given plant is left invertible and has all its invariant zeros in the left half s plane (i.e. of minimum phase).

Moreover, a constructive method of obtaining a gain K to achieve both closed-loop stability and

ELTRI can be given under the above three conditions. Such a gain K in general is nonunique and belongs to a class of gains denoted by \mathcal{K}_r .

Proof. Under the conditions given in the theorem, a method of calculating the class of gains \mathcal{K}_r is given in Appendix A.

Remark 1. There is no method whatsoever in the literature to obtain the observer gain K that achieves ELTRI. Although this paper is not intended in general to give methods of obtaining K , the constructive proof of Theorem 1 yields one such method.

Remark 2. It is important to realize the implications of the condition $FB = 0$. Apparently, it restricts the class of loop transfer functions $L(s)$ that are attainable by full state feedback. In particular, under the condition $FB = 0$,

$$L(s) = F\Phi B \cdots \frac{FAB}{s^2},$$

implying that $\|L(j\omega)\|$ must have at least a roll-off of 40 dB per decade with respect to ω . It is well known that whenever the state feedback gain F is calculated by LQR theory, $\|L(j\omega)\|$ has only a roll-off of 20 dB per decade with respect to ω . Thus the use of LQR theory is then ruled out to generate the target loop transfer function.

Since $FB = 0$ severely restricts the class of loop transfer functions that are achievable, most of the existing literature focuses attention on ALTRI methods. In these ALTRI methods, one tries to find a gain K such that (1.13) is satisfied. As we discussed earlier, the gain K in this case is parameterized in terms of a tuning parameter σ . Satisfying (1.13) is a sufficient condition to render $\sigma_{\max}[M(j\omega)]$ arbitrarily small for all ω . At first, Doyle and Stein (1979) gave a sufficient condition under which (1.13) is true. Their condition is as follows: Let $K(\sigma)$ be chosen such that as $\sigma \rightarrow \infty$, $K(\sigma)/\sigma \rightarrow BW$ for some non-singular matrix W . Then, (1.13) is true and consequently ALTRI is achieved as $\sigma \rightarrow \infty$. There were several attempts later on to weaken the Doyle-Stein condition (Madiwale and Williams, 1985; Matson and Maybeck, 1987; Saberi and Sannuti, 1988). It is well known that in order to satisfy the Doyle-Stein condition, one requires only that the plant be left invertible and be of minimum phase. Thus in comparison with the sufficient conditions for ELTRI as stated in Theorem 1, one finds a drastic relaxation of the required conditions for ALTRI.

As far as the design of $K(\sigma)$ is concerned, presently there exists three different methods: (1) asymptotic LQG methods, (2) asymptotic pole placement methods and (3) eigenstructure assignment methods. An exhaustive comparison of all these three methods is given in Saberi and Sannuti (1988). All these procedures aim at obtaining a gain K such that (a) some of the observer eigenvalues either coincide or are close to the zeros of the given plant and that the associated left eigenvectors satisfy the condition $V_i^H B = 0$ either exactly or approximately, and (b) the remaining observer eigenvalues are placed far in the left half s plane so that the corresponding $R_i/(s - \lambda_i)$ in the dyadic expansion (2.7) are approximately zero. In other words, all these methods find a gain K such that $\sigma_{\max}[M(j\omega)]$ is arbitrarily small for all ω . Thus the term $M(s)$ plays a dominant role in LTR. The following result summarizes this discussion.

Theorem 2. Consider the closed-loop system comprising of the plant and the full order observer based controller as in Fig. 2. Let the given plant be left invertible and be of minimum phase. Then a gain $K(\sigma)$ can be designed such that both asymptotic stability of the closed-loop system and ALTRI can be achieved. Such a gain $K(\sigma)$ in general is nonunique and belongs to a class of gains denoted by $\mathcal{H}_\sigma(\sigma)$.

Proof. See Saberi and Sannuti (1988).

Let us next examine the eigenvalues of the observer based controller. These eigenvalues are given by

$$\lambda(A - KC - BF)$$

These eigenvalues are not necessarily in the left half s plane for all K . To study the nature of these eigenvalues, consider the following:

$$\begin{aligned} \det[sI_n - A + KC + BF] &= \det[sI_n - A + KC] \\ &\quad \times \det[I_n + (\Phi^{-1} + KC)^{-1}BF] \\ &= \det[sI_n - A + KC] \\ &\quad \times \det[I_m + F(\Phi^{-1} + KC)^{-1}B] \\ &= \det[sI_n - A + KC] \\ &\quad \times \det[I_m + M(s)]. \end{aligned} \quad (2.8)$$

Thus whenever ELTRI is achieved, i.e. whenever $M(s) = 0$, the controller eigenvalues are given by $\lambda(A - KC)$. Hence the observer-based controller is asymptotically stable. On the other hand, in the case of ALTRI, the eigenvalues of the full order observer-based controller obviously approach $\lambda(A - KC)$ as $\sigma \rightarrow \infty$. How-

ever, in practice the value of σ needed for the desired degree of recovery might not yield an asymptotically stable controller. In fact, this is the case in most practical problems.

As discussed above, most often one opts for ALTRI design as it requires less stringent conditions than ELTRI design. In ALTRI, the level of recovery depends on $\sigma_{\max}[M(j\omega)]$. However in order to render $\sigma_{\max}[M(j\omega)]$ small, one needs to increase the tuning parameter σ which itself increases the gain $K(\sigma)$. Thus as discussed thoroughly by Sogaard-Andersen and Niemann (1989), there is a fundamental trade-off between the level of recovery and the size of gain. This trade-off can be visualized in a natural way in terms of the trade-off between the singular values of sensitivity and complementary sensitivity functions and the singular values of $M(j\omega)$. The reason for this is that the robust stability and nominal performance of a system are directly reflected in the singular values of sensitivity and complementary sensitivity functions, whereas the level of recovery (i.e. the size of $F_0(j\omega)$) is directly dependent on the singular values of $M(j\omega)$. With this point in view, Sogaard-Andersen and Niemann (1989) derive some analytical expressions for the discrepancy between the desired and the achieved sensitivity and complementary sensitivity functions. Let $S_i(s)$ and $I_0(s)$ be the achieved sensitivity and complementary sensitivity functions in the configuration of Fig. 2 when the loop is broken at the input point of the plant

$$S_0(s) = [I_m + C_0(s)P(s)]^{-1}$$

and

$$I_0(s) = I_m - S_0(s) = [I_m + C_0(s)P(s)]^{-1}C_0(s)P(s)$$

where $C_0(s)$ is as given in (1.6). Let $S_f(s)$ and $I_f(s)$ be the sensitivity and complementary sensitivity functions corresponding to the target loop-shape. The following lemma is a slight generalization of the results of Sogaard-Andersen and Niemann (1989).

Lemma 5. Consider the configuration of Fig. 2. We have the following bounds on all singular values $i = 1$ to m of $S_0(j\omega)$ and $I_0(j\omega)$:

$$\frac{|\sigma_i[S_0(j\omega)] - \sigma_i[S_f(j\omega)]|}{\sigma_{\max}[S_f(j\omega)]} \leq \sigma_{\max}[M(j\omega)],$$

and

$$\frac{|\sigma_i[I_0(j\omega)] - \sigma_i[I_f(j\omega)]|}{\sigma_{\max}[I_f(j\omega)]} \leq \sigma_{\max}[M(j\omega)].$$

The expressions given above can be used to analyze the inevitable trade-off between good

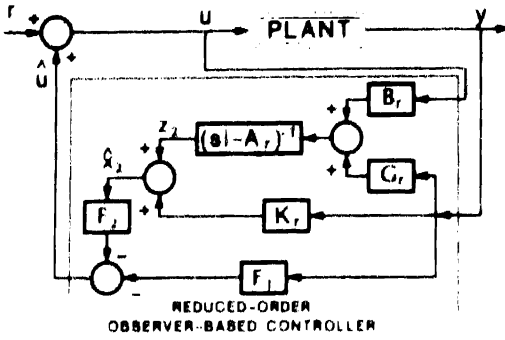


FIG. 4. Plant with reduced order observer based controller.

recovery as indicated by $\sigma_{\max}[M(j\omega)]$ and robustness and performance as reflected in the sensitivity and complementary sensitivity functions. To do this, Sogaard-Andersen and Niemann (1989) developed some recovery diagrams.

Before closing this section, let us note that $M(s)$ plays a central role in every single result given in this section.

2.2. Reduced order observers—ELTRI and ALTRI

Now let us consider a reduced order observer based controller as in Fig. 4. Without loss of generality, let us assume that

$$C = [I_p, 0]$$

and hence the plant (1.1) is in the form,

$$\dot{x}_1 = A_{11}x_1 + A_{12}x_2 + B_1u \quad (2.9)$$

$$\dot{x}_2 = A_{21}x_1 + A_{22}x_2 + B_2u, \quad (2.10)$$

$$y = x_1.$$

Also, let the state feedback gain matrix F which achieves the target loop transfer function $L(s)$ be partitioned in conformity with (2.9) as

$$F = [F_1, F_2]. \quad (2.11)$$

Let $\Phi_{11} = (sI_p - A_{11})^{-1}$ and $\Phi_{22} = (sI_{n-p} - A_{22})^{-1}$. It is then straightforward to derive the following relationships:

$$P(s) = C\Phi B = \Phi_{11}B_1 + \Phi_{11}A_{12}H_2(s), \quad (2.12)$$

$$L(s) = F\Phi B = F_1P(s) + F_2H_2(s), \quad (2.13)$$

where $H_2(s)$ is the transfer function between u and x_2 , i.e. $x_2(s) = H_2(s)u(s)$, where

$$H_2(s) = (\Phi_{22}^{-1} - A_{21}\Phi_{11}A_{12})^{-1}(A_{21}\Phi_{11}B_1 + B_2). \quad (2.14)$$

The reduced order observer equations (O'Reilly, 1983; Madiwale and Williams, 1985) are given by

$$\dot{z}_2 = A_r z_2 + G_r y + B_r u, \quad (2.15)$$

with

$$x_1 = y \quad \text{and} \quad \hat{x}_2 = K_r y + z_2, \quad (2.16)$$

and with the reduced order observer based feedback control law as

$$u = \hat{u} = -F_1 x_1 - F_2 \hat{x}_2. \quad (2.17)$$

Here K_r is the reduced order observer gain and the matrices A_r , B_r and G_r are given by

$$\begin{aligned} A_r &= A_{22} - K_r A_{12}, \quad B_r = B_2 - K_r B_1, \\ G_r &= A_{22} K_r - K_r A_{12} K_r + A_{21} - K_r A_{11}. \end{aligned} \quad (2.18)$$

Now in order to bring the theory of full and reduced order observers to the same frame work and to understand the conditions for either ELTRI or ALTRI clearly, we present the following results which are analogous to Lemmas 1 and 2.

Lemma 6. $E_{or}(s)$, the error between the target loop transfer function $L(s)$ and that realized by the reduced order observer-based controller of Fig. 4 is given by

$$E_{or}(s) = M_r(s)(I_m + M_r(s))^{-1}(I_m + F\Phi B), \quad (2.19)$$

where

$$M_r(s) = F_2(\Phi_{22}^{-1} + K_r A_{12})^{-1} B_r. \quad (2.20)$$

Proof. See Appendix B.

Remark 3. The expression for $E_{or}(s)$ is identical to the corresponding one when full order observer-based controller is used; see (2.4), except that now $M_r(s)$ takes the place of $M(s)$.

Lemma 7.

$$E_{or}(j\omega) = 0 \quad \text{iff} \quad M_r(j\omega) = 0 \quad \text{for all} \quad \omega \in \Omega_r, \quad (2.21)$$

where Ω_r is the set of all $0 \leq \omega < \infty$ for which $L_{or}(j\omega)$ and $L(j\omega)$ are well defined (i.e. all required inverses exist).

Proof. The proof is obvious in view of Lemma

As in the case of a full order observer, a physical interpretation can be given to the term $M_r(s)$. It is straightforward to show that

$$-\hat{u}(s) = M_r(s)u(s) + [F_1 + F_2(\Phi_{22}^{-1} + K_r A_{12})^{-1} G_r + K_r(s)]y(s).$$

Thus we note that $M_r(s)$ is the transfer function from u to $-\hat{u}$. Hence as in the case of a full order observer, whenever the size of $M_r(s)$ is small, the effect of the link from the input point

of the plant to the observer, on \hat{u} is small. In view of Lemma 7, the question now is when and how $M_r(j\omega)$ can be made either exactly or approximately zero for all ω . We note that,

$$B_r = B_2 - K_r B_1 = 0, \quad (2.22)$$

is a sufficient condition for $\{M_r(j\omega)\}$ to be identically zero. Unlike in full order observers, the condition (2.22) involves K_r and hence there is a possibility of solving for K_r from it. Sogaard-Andersen (1987b) under the conditions that (a) the given system is square and invertible and (b) B_1 is nonsingular, solves for K_r ,

$$K_r = B_2 B_1^{-1} \quad (2.23)$$

It turns out that a gain K_r which satisfies (2.22) and thus achieves ELTR can be obtained under much relaxed conditions. We have the following result analogous to Theorem 1.

Theorem 3. Consider the closed-loop system comprising of the plant and the reduced order observer based controller as in Fig. 4. Then both the asymptotic stability of the closed-loop system and ELTRI can be achieved under the following conditions:

1. The given plant has all its infinite zeros of order one.
2. The given plant is left invertible and is of minimum phase.

Moreover, a constructive method of obtaining a gain K_r to achieve both ELTRI and asymptotic stability of the closed-loop system can be given under the above two conditions. Such a gain K_r in general is nonunique and belongs to a class of gains denoted by \mathcal{K}_{er} .

Proof. Under the conditions given in the theorem, a method of calculating the class of gains \mathcal{K}_{er} is given in Appendix C.

Remark 4. When reduced order observer based controllers are used, the condition $FB = 0$ is not necessary. However, while full order observer always results in a strictly proper controller transfer function, the reduced order observer based controller has a nonstrictly proper transfer function. As discussed by Khalil (1981, 1984) and by Vidyasagar (1985a), a closed-loop system with a nonstrictly proper controller is not robust under unmodelled high frequency dynamics.

Remark 5. Madiwale and Williams (1985) gave a sufficient condition for ELTRI,

$$K_r(I_p + A_{12}\Phi_{22}K_r)^{-1}A_{12}\Phi_{22}B_r = B_r. \quad (2.24)$$

There is a one to one correspondence between

(2.2) and (2.24). In the same way as we showed (2.3) is equivalent to $B = 0$, we can show that (2.24) is equivalent to $B_r = 0$.

Since in general $M_r(s)$ cannot exactly be made zero, one focuses attention on ALTRI. That is, one needs

$$M_r(s) = F_2(\Phi_{22}^{-1} + K_r(\sigma)A_{12})^{-1}B_r \rightarrow 0$$

pointwise in s as $\sigma \rightarrow \infty$,

where the gain $K_r(\sigma)$ is now parameterized in terms of a tuning parameter σ . However, as in the previous section, in order to have the state feedback and observer designs to be independent of one another, one needs to require that

$$(\Phi_{22}^{-1} + K_r(\sigma)A_{12})^{-1}B_r \rightarrow 0 \text{ pointwise in } s \text{ as } \sigma \rightarrow \infty. \quad (2.25)$$

To design such a $K_r(\sigma)$, Dowdle *et al.* (1982) study a restrictive class of systems where all the first Markov parameters of the given plant are zero. Such a severe restriction on the given plant is not imposed in Madiwale and Williams (1985), instead they require that certain matrices are of full rank and a certain subsystem of the given system is of minimum phase. Based on asymptotic LQG methods, Sogaard-Andersen (1985b) studies the general case without any restrictions. Since Sogaard-Andersen divides it into three different cases, his analysis and design besides being not unified becomes unnecessarily involved. Saberi and Sannuti (1988) give an explicit method of calculating the gain $K_r(\sigma)$ which satisfies the condition (2.25). In fact, they convert the problem of designing the reduced order observer for the given plant into that of a full order observer, however, for a reduced order subsystem of the given plant.

The above discussion can be summarized as Theorem 4 which is analogous to Theorem 2.

Theorem 4. Consider the closed-loop system comprising of the plant and the reduced order observer based controller as in Fig. 4. Let the given plant be left invertible and be of minimum phase. Then a gain $K_r(\sigma)$ can be designed such that both asymptotic stability of the closed-loop system and ALTRI can be achieved. Such a gain $K_r(\sigma)$ in general is nonunique and belongs to a class of gains denoted by $\mathcal{K}_{er}(\sigma)$.

Proof. See Saberi and Sannuti (1988).

Let us next examine the eigenvalues of the reduced order observer based controller. These eigenvalues are given by

$$\lambda(A_{22} - K_r A_{12} - B_2 F_2).$$

These eigenvalues are not necessarily in the left half s plane for all K . As in (2.8), we can show that

$$\begin{aligned} \det [sI_{n-p} - A_{22} + K_r A_{12} + B_2 F_2] \\ = \det [sI_{n-p} - A_{22} + K_r A_{12}] \det [I_m + M_r(s)]. \end{aligned} \quad (2.26)$$

Thus whenever ELTRI is achieved, i.e. whenever $M_r(s) = 0$, the controller eigenvalues are given by $\lambda(A_r)$ and hence the reduced order observer based controller is asymptotically stable. On the other hand in the case of ALTRI, the eigenvalues of the reduced order observer based controller tend to $\lambda(A_r)$ as $\sigma \rightarrow \infty$. However, as in full order observers, the value of σ needed for the desired degree of recovery might not yield an asymptotically stable controller. In fact, this is the case in most practical problems.

Now as in Lemma 5, we would like to develop bounds on the sensitivity and complementary sensitivity functions generated by the use of reduced order observer based controllers. Let $S_{or}(s)$ and $T_{or}(s)$ be the generated sensitivity and complementary sensitivity functions in the configuration of Fig. 4 when the loop is broken at the input point of the plant,

$$S_{or}(s) = [I_m + C_{or}(s)P(s)]^{-1}$$

and

$$\begin{aligned} T_{or}(s) &= I_m - S_{or}(s) \\ &= [I_m + C_{or}(s)P(s)]^{-1} C_{or}(s)P(s) \end{aligned}$$

where $C_{or}(s)$ is the transfer function of the reduced order observer based controller. We have the following result analogous to Lemma 5.

Lemma 8. Consider the configuration of Fig. 4. We have the following bounds on all singular values $i = 1$ to m of $S_{or}(j\omega)$ and $T_{or}(j\omega)$:

$$\frac{|\sigma_i[S_{or}(j\omega)] - \sigma_i[S_F(j\omega)]|}{\sigma_{\max}[S_F(j\omega)]} \leq \sigma_{\max}[M_r(j\omega)],$$

and

$$\frac{|\sigma_i[T_{or}(j\omega)] - \sigma_i[T_F(j\omega)]|}{\sigma_{\max}[S_F(j\omega)]} \leq \sigma_{\max}[M_r(j\omega)].$$

Proof. See Appendix D.

2.3. Full and reduced order observers—ELTRO and ALTRO

The target open-loop transfer functions can be designed when the loop is broken at either the input or the output point of the given plant depending upon the given specifications. We have discussed so far LTR recovery at the input

point, either exact or approximate type (ELTRI or ALTRI), using either full or reduced order observer-based controllers. Now we would like to consider LTR recovery when the loop is broken at the output point (Kwakernaak, 1969). This method is used when the designer specifications and the modelling of uncertainties are reflected at the output point of the plant. In the literature, it is commonly said that LTR recovery at the input and output points (LTRI and LTRO) are dual to one another. This duality is well understood in the case when full order observers or Kalman filters are used in the controllers. That is, in the case of LTRO, the first step is to design a Kalman filter, via loop shaping techniques, whose loop transfer function meets the design specifications. The next step is to recover this Kalman filter loop transfer function via LTR technique. However, this kind of duality is not well understood when reduced order observer based controllers are used. For instance, Sogaard-Andersen (1987b) who has contributed much to the development of reduced or minimal order observers for LTRI makes the following comment: "The loop-shape formulation used here requires that the uncertainties and performance specifications are reflected to the plant input. Unfortunately similar results for the plant output cannot be derived since the minimal-order observer and the plant model are not dual." The confusion arises here because the duality is sought between the plant and the observer. The proper way is to seek the duality in the design methodology and when this is done, contrary to the statement of Sogaard-Andersen, minimal order observer based controllers can be designed for LTRO as well. In other words, one needs first to clearly define the duality in a mathematical way and then needs to interpret the implications of it as to the controller implementation. In order to avoid any confusion, we give below a formal step by step algorithm to show how duality arises for LTR recovery at the input and output points.

1. Let the given plant model Σ be characterized by the triple (A, B, C) where A , B and C are respectively $n \times n$, $n \times m$ and $p \times n$ matrices. Let Σ be of minimum phase and be right invertible implying $p \leq m$. Also, let $P(s)$ be the transfer function of the plant Σ ,

$$P(s) = C(sI_n - A)^{-1}B.$$

Let $L(s)$ be the required target open-loop transfer function when the loop is broken at the output point of the given plant. Thus, in the configuration of Fig. 1, we are seeking a

controller $C(s)$ such that $L(j\omega)$ is either exactly or approximately equal to $P(j\omega)C(j\omega)$.

2. Define a transposed system model Σ_r characterized by the triple (A_r, B_r, C_r) where

$$A_r \equiv A', \quad B_r \equiv C', \quad C_r \equiv B'.$$

Note that since Σ is of minimum phase and right invertible, Σ_r is of minimum phase and left invertible. Also, note that $P_r(s)$, the transfer function of the plant Σ_r is $P'(s)$. Let $L_r(s)$ be defined as

$$L_r(s) \equiv L'(s).$$

3. For the purpose of design alone, consider the fictitious plant Σ_r as given in step 2. Then design a controller $C_r(s)$ such that $C_r(j\omega)P_r(j\omega)$ is either exactly or approximately equal to $L_r(j\omega)$. For this purpose one can use either a full or reduced order observer based controller design of Sections 2.1 or 2.2. In fact one can also use any other compensator design schemes such as those to be described in Sections 3 and 4. We note that the dynamic order of $C_r(s)$ is either n or $n - m$ depending upon either full or reduced order observer is used for the controller design.

4. Define a controller $C(s)$:

$$C(s) \equiv C_r'(s).$$

We note that the dynamic order of $C(s)$ is same as that of $C_r(s)$.

Then it can be shown trivially that the controller $C(s)$ designed above and implemented as in Fig. 1 achieves either ELTRO or ALTRO depending upon whether $C_r(s)$ in step 3 is designed to achieve ELTRI or ALTRI for the fictitious plant Σ_r .

3. FULL ORDER COMPENSATOR: ELTRI AND ALTRI

In this section and the next, we present our main contributions. To start with, let us recall the concept of LTRI. Given the plant transfer function $P(s) \equiv C\Phi B$ and the target loop transfer function $L(s) \equiv F\Phi B$, one wants to design a controller with transfer function $C(s)$ such that $C(j\omega)P(j\omega)$ is either exactly or approximately equal to $L(j\omega)$. The only controller that is available so far for this purpose is observer based. As reviewed in the last section, in observer based controllers, $M(s)$ plays a central role in the recovery procedure. Numerical experience shows that in order to achieve a satisfactory degree of recovery, large values of gain in general are required by the observer based controllers. In an attempt to reduce the size of required gains, one then naturally seeks new structures for the control-

lers. The physical meaning of the transfer function $M(s)$ as explained in the last section, leads us to examine the observer based controller structure in which the link from the input point of the plant via the control distribution matrix B to the controller is removed. Such an omission of the link generates a new structure for the controller which we now call a compensator. Because of the omission of the link mentioned earlier, the celebrated separation principle is no longer valid and hence the properties of the compensator as to closed-loop stability and achieving LTR have to be examined carefully. This is the purpose of this section and the next.

Consider the dynamic compensator,

$$(A - KC)z + Kv, \quad (3.1)$$

$$u \equiv u - Fz. \quad (3.2)$$

The only unknown matrix in (3.1) is K which is considered as a free design parameter. The compensator transfer function (i.e., the transfer function from y to $-u$) is given by

$$C_r(s) \equiv F(\Phi^{-1} + KC)^{-1}K. \quad (3.3)$$

We would like to design K to satisfy the following conditions:

1. *Stability of the closed-loop system.* The closed-loop system as depicted in Fig. 3 and characterized by (1.1), (3.1) and (3.2), is asymptotically stable, i.e.

$$\text{Re}[\lambda(A_r)] < 0, \quad (3.4)$$

where

$$\begin{aligned} A - KC &= KC \\ BF &= A \end{aligned} \quad (3.5)$$

2. *ELTRI or ALTRI.* The achieved loop transfer function $L_r(j\omega)$,

$$L_r(j\omega) \equiv C_r(j\omega)P(j\omega), \quad (3.6)$$

is either exactly or approximately equal to $L(j\omega)$.

3. *Open-loop stability of the compensator.* The compensator is open-loop asymptotically stable, i.e.

$$\text{Re}[\lambda(A - KC)] < 0. \quad (3.7)$$

The above three conditions are important from technical point of view. However, merely determining K to satisfy the above conditions is not enough because our primary goal as stated earlier is to come up with a scheme which requires smaller values of gain than the observer based controller to achieve the same level of LTR. In what follows, we show that our new compensator structure does exactly this. We first give the following lemma analogous to Lemma 1.

Lemma 9. $E_e(s)$, the error between the target open-loop transfer function $L(s)$ and $L_r(s)$, the one realized by the compensator, is given by

$$E_e(s) = M(s), \quad (3.8)$$

where

$$M(s) = F(\Phi^{-1} + KC)^{-1}B. \quad (3.9)$$

Proof.

$$\begin{aligned} E_e(s) &= L(s) - L_r(s) \\ &= F[I_n - (\Phi^{-1} + KC)^{-1}KC]\Phi B \\ &= F(\Phi^{-1} + KC)^{-1}B. \end{aligned}$$

Remark 6. Observe that $M(s)$ defined here is exactly the same as the one defined earlier for the full order observer-based controllers, see (2.5). In view of this, the two expressions for the error between the required and the achieved loop transfer functions, one for the conventional observer-based design (2.4) and the other for the new compensator design (3.8), differ significantly. This as we shall see later on in Theorems 7 and 8 leads to an overwhelming advantage in favor of the new compensator approach.

Since $M(s)$ defined here and in the case of full order observer based design is one and the same, we naturally see that ELTRI or ALTRI is achievable by the new compensator under exactly the same conditions as in the previous case. That is K has to be an element of either \mathcal{K}_r for ELTRI or an element of $\mathcal{K}_a(\sigma)$ for ALTRI. However, the stability of the closed-loop system has to be separately examined. We have the following theorem.

Theorem 5. Consider the closed-loop system comprising of the plant along with the compensator as in Fig. 3. Then both the asymptotic stability of the closed-loop system and ELTRI can be achieved under the following conditions:

1. $FB = 0$.
2. The given plant has all its infinite zeros of order one.
3. The given plant is left invertible and is of minimum phase.

Moreover, under the above conditions, K can be selected as an element of \mathcal{K}_r . Also, the eigenvalues of A_{cl} are given by $\lambda(A - KC)$ and $\lambda(A - BF)$. Furthermore, the developed compensator is always open-loop asymptotically stable.

Proof. Under the conditions given and in view of the Theorem 1, it is obvious that K can be

selected as an element of \mathcal{K}_r and hence $M(s) = 0$. It is also evident that the compensator is open-loop asymptotically stable. Next, the closed-loop stability can be proved as follows. The dynamic matrix of the closed-loop system is given by (3.5). Then consider the following reductions:

$$\begin{aligned} \det[sI_{2n} - A_{cl}] &= \det \begin{bmatrix} sI_n - A + KC & -KC \\ BF & sI_n - A \end{bmatrix} \\ &= \det \begin{bmatrix} \Phi^{-1} & -KC \\ \Phi^{-1} + BF & \Phi^{-1} \end{bmatrix} \\ &= \det \begin{bmatrix} \Phi^{-1} & -KC \\ BF & \Phi^{-1} + KC \end{bmatrix} \\ &= \det \begin{bmatrix} \Phi^{-1} + BF & \Phi^{-1} \\ BF & \Phi^{-1} + KC \end{bmatrix}. \end{aligned} \quad (3.10)$$

Now using Schur's formula for calculating the determinant of a partitioned matrix, we have

$$\begin{aligned} \det[sI_{2n} - A_{cl}] &= \det[\Phi^{-1} + KC] \\ &\quad \times \det[\Phi^{-1} + BF - \Phi^{-1}(\Phi^{-1} + KC)^{-1}BF] \\ &= \det[\Phi^{-1} + KC] \det[\Phi^{-1}] \\ &\quad \times \det[I_n + \Phi BF - (\Phi^{-1} + KC)^{-1}BF] \\ &= \det[\Phi^{-1} + KC] \det[\Phi^{-1}] \\ &\quad \times \det\{I_n + [\Phi B - (\Phi^{-1} + KC)^{-1}B]F\}. \end{aligned} \quad (3.11)$$

Now using the identity

$$\det[I_n + A_1 A_2] = \det[I_m + A_2 A_1] \quad (3.12)$$

for any $n \times m$ and $m \times n$ matrices A_1 and A_2 ,

$$\begin{aligned} \det[sI_{2n} - A_{cl}] &= \det[\Phi^{-1} + KC] \det[\Phi^{-1}] \\ &\quad \times \det\{I_m + F[\Phi B - (\Phi^{-1} + KC)^{-1}B]\} \\ &= \det[\Phi^{-1} + KC] \det[\Phi^{-1}] \\ &\quad \times \det\{I_m + F\Phi B - F(\Phi^{-1} + KC)^{-1}B\} \\ &= \det[\Phi^{-1} + KC] \det[\Phi^{-1}] \\ &\quad \times \det\{I_m + F\Phi B - M(s)\}. \end{aligned} \quad (3.13)$$

Noting that $M(s) = 0$, (3.13) reduces to

$$\begin{aligned} \det[sI_{2n} - A_{cl}] &= \det[\Phi^{-1} + KC] \det[\Phi^{-1}] \det\{I_m + F\Phi B\} \\ &= \det[\Phi^{-1} + KC] \det[\Phi^{-1} + BF]. \end{aligned}$$

Since by design $A - KC$ and $A - BF$ are asymptotically stable matrices, the closed-loop system of Fig. 3 is then asymptotically stable.

As in the case of full order observer based controllers, the conditions given in Theorem 5, especially the conditions 1 and 2 are very restrictive and hence are not true for many

practical systems. To broaden the class of systems, one abandons the goal of achieving ELTRI and instead seeks ALTRI. For this purpose, as in the previous section, we parameterize K in terms of a tuning parameter σ . We have the following theorem dealing with ALTRI.

Theorem 6. Consider the closed-loop system comprising of the plant along with the compensator as in Fig. 3. Assume that the given plant is left invertible and is of minimum phase. Select the gain K which is parameterized in terms of a tuning parameter σ , as an element of $\mathcal{K}_s(\sigma)$. Then ALTRI is achieved as $\sigma \rightarrow \infty$. Furthermore, there exists a σ_1 such that the closed-loop system is asymptotically stable for all $\sigma > \sigma_1$. More specifically, as $\sigma \rightarrow \infty$, eigenvalues of A_{cl} are given by

$$\lambda(A - K(\sigma)C) + O(1/\sigma)$$

and

$$\lambda(A - BF) + O(1/\sigma).$$

Also, the developed compensator is always open-loop asymptotically stable.

Proof. The results of achieving ALTRI and open-loop asymptotic stability of the compensator are obvious. The proof of closed-loop stability of Fig. 3 can be seen as follows. In view of (3.13) and noting that $M(s)$ tends to zero point wise in s as $\sigma \rightarrow \infty$, we have

$$\begin{aligned} \det[sI_{2n} - A_{cl}] &\rightarrow \det[\Phi^{-1} + K(\sigma)C] \det[\Phi^{-1}] \\ &\quad \times \det[I_n + \Phi BF] \quad \text{as } \sigma \rightarrow \infty \\ &= \det[\Phi^{-1} + K(\sigma)C] \det[\Phi^{-1} + BF]. \end{aligned} \quad (3.14)$$

This completes the proof of Theorem 6.

Remark 7. The full order observer is not in general open-loop stable while open-loop stability of the compensator is always guaranteed.

As discussed earlier, one often opts for ALTRI design as it requires less stringent conditions than ELTRI design. However, ALTRI is fundamentally an asymptotic result. In practice, the degree of recovery depends on the size of gain. Both conventional observer-based controller and our new compensator are capable of achieving ALTRI. The following theorem however shows that for the same value of gain, the new compensator achieves much better degree of recovery than the observer-based controller.

Theorem 7. Let $K(\sigma)$ be an element of $\mathcal{K}_s(\sigma)$. Assume also that the same gain $K(\sigma)$ is used for

both the observer-based controller and for the new compensator. Let σ be such that $\sigma_{\max}[M(j\omega)]$ is small (say, $\ll 1$) for all ω . Furthermore, assume that

$$\begin{aligned} \sigma_{\min}[L(j\omega)] &= \sigma_{\min}[F(j\omega - A)^{-1}B] \gg 1 \quad \text{for all } \omega \in D_i, \end{aligned} \quad (3.15)$$

for some frequency region of interest, D_i . Then for all $\omega \in D_i$, the mismatch between the target loop transfer function and the one achieved by the compensator is always less than the corresponding one achieved by the full order observer-based controller. More specifically, we have

$$\sigma_{\max}[E_0(j\omega)] \gg \sigma_{\max}[E_c(j\omega)] \quad \text{for all } \omega \in D_i, \quad (3.16)$$

where $E_c(s)$ is as in (3.8) and $E_0(s)$ is as in (2.4).

Proof. Recalling the expression for $E_0(j\omega)$ from (2.4), we have

$$\begin{aligned} \sigma_{\max}[E_0(j\omega)] &\approx \sigma_{\max}\{M(j\omega)[I_m + M(j\omega)]^{-1}(I_m + F\Phi(j\omega)B)\} \\ &\approx \sigma_{\max}[M(j\omega)]\sigma_{\min}\{|I_m + M(j\omega)|^{-1}\} \\ &\quad \times \sigma_{\min}[I_m + F\Phi(j\omega)B] \\ &= \frac{\sigma_{\max}[M(j\omega)]\sigma_{\min}[I_m + F\Phi(j\omega)B]}{\sigma_{\max}[I_m + M(j\omega)]} \\ &\approx \sigma_{\max}[E_c(j\omega)]\alpha(j\omega), \end{aligned} \quad (3.17)$$

where

$$\alpha(j\omega) = \frac{\sigma_{\min}[F\Phi(j\omega)B] - 1}{1 + \sigma_{\max}[M(j\omega)]}$$

Now by our assumption, $\sigma_{\max}[M(j\omega)]$ is $\ll 1$ and $\sigma_{\min}[F\Phi(j\omega)B]$ is $\gg 1$ for all $\omega \in D_i$, and hence $\alpha(j\omega)$ is $\gg 1$ for all $\omega \in D_i$. Thus

$$\sigma_{\max}[E_0(j\omega)] \gg \sigma_{\max}[E_c(j\omega)] \quad \text{for all } \omega \in D_i.$$

Remark 8. It is well known (Doyle and Stein, 1981) that in order to have good command following and disturbance rejection properties, the loop transfer function matrix $L(j\omega)$ has to be large and consequently, the minimum singular value $\sigma_{\min}[L(j\omega)]$ should be large in the appropriate frequency region. Thus the condition (3.15) is always satisfied in all practical situations.

Remark 9. Theorem 7 is intuitively evident. In our compensator $E_c(s)$, the error between the required and the achieved loop transfer function is equal to $M(s)$ which is designed to be small in some sense. On the other hand, in conventional observer-based design, the corresponding error

$E_0(s)$, is a multiple of $M(s)$, $[I_m + M(s)]^{-1}$ and $I_m + F\Phi B$. But in any good design, the loop transfer function $F\Phi B$ is large in the frequency region of interest. Thus for the same gain $K(\sigma)$, $E_0(s)$ differs from $E_c(s)$ by a large factor ($\approx \|F\Phi B\|$) making $E_0(s)$ much worse than $E_c(s)$.

Once again, as in Lemma 5, we now develop bounds on sensitivity and complementary sensitivity functions when the new compensator is used. Let $S_c(s)$ and $T_c(s)$ be the generated sensitivity and complementary sensitivity functions in the configuration of Fig. 3 when the loop is broken at the input point of the plant,

$$S_c(s) = [I_m + C_c(s)P(s)]$$

and

$$T_c(s) = I_m - S_c(s) = [I_m + C_c(s)P(s)]^{-1}C_c(s)P(s)$$

where $C_c(s)$ is as in (3.3). We have the following result analogous to Lemma 5.

Theorem 8. Consider the configuration of Fig. 3. Assume that (3.15) is true. We have the following bounds on all singular values $i = 1$ to m of $S_c(j\omega)$ and $T_c(j\omega)$:

$$\begin{aligned} & \frac{\sigma_i[S_c(j\omega)] - \sigma_i[S_r(j\omega)]}{\sigma_{\max}[S_r(j\omega)]} \\ & \leq \frac{\sigma_{\max}[M(j\omega)]}{\sigma_{\min}[F\Phi(j\omega)B] - \sigma_{\max}[M(j\omega)] - 1} \\ & \ll \sigma_{\max}[M(j\omega)] \quad \text{for all } \omega \in D_c \end{aligned} \quad (3.18)$$

and

$$\begin{aligned} & \frac{\sigma_i[T_c(j\omega)] - \sigma_i[T_r(j\omega)]}{\sigma_{\max}[S_r(j\omega)]} \\ & \leq \frac{\sigma_{\max}[M(j\omega)]}{\sigma_{\min}[F\Phi(j\omega)B] - \sigma_{\max}[m(j\omega)] - 1} \\ & \ll \sigma_{\max}[M(j\omega)] \quad \text{for all } \omega \in D_c \end{aligned} \quad (3.19)$$

Proof. See Appendix E.

Remark 10. It is evident that due to the presence of the sign \ll in the expressions (3.18) and (3.19), the new compensator yields much better sensitivity and complimentary sensitivity recovery than the conventional full order observer based controller.

4. REDUCED ORDER COMPENSATOR—ELTRI AND ALTRI

In the previous section, we studied a new compensator whose dynamic order is the same as that of the given plant. Corresponding to the reduced order observer-based controllers, one naturally would like to investigate also the

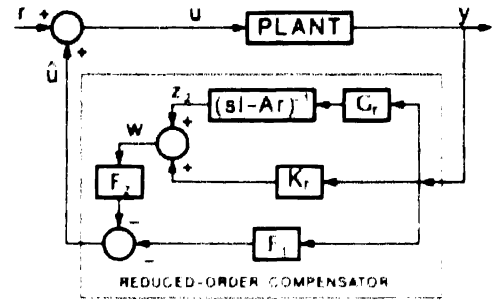


FIG. 5. Plant with reduced order compensator

possibility of a reduced order compensator for the purpose of achieving either ELTRI or ALTRI. Motivated by the results of the previous section, in this section we study such a reduced order compensator structure (see Fig. 5).

The structure shown in Fig. 5 corresponds to that of the reduced order observer-based controller except that, as in the case of full order compensator, the link from the plant input point via the matrix B to the controller (or what is now called a compensator) is omitted. Because of this omission of the link, again the separation principle is no longer valid and hence we need to study and establish the necessary properties of the reduced order compensator for LTRI.

As in Section 2, without loss of generality, we will assume that the given plant is described by (2.9) and (2.10) while the state feedback gain matrix F which achieves the target loop transfer function $L(s)$ be partitioned as in (2.11). The reduced order compensator is dynamically described by

$$\dot{z}_2 = A_r z_2 + G_r y, \quad (4.1)$$

$$u = \hat{u} = -F_1 x_1 - F_2 w, \quad (4.2)$$

$$w = K_r y + z_2. \quad (4.3)$$

The matrices A_r and G_r are as in (2.18). Here K_r is a free design parameter which is to be selected to satisfy the following conditions:

1. *Stability of the closed-loop system.* The closed-loop system as depicted in Fig. 5 and characterized by (2.9), (2.18), (4.1) to (4.3), is asymptotically stable, i.e.

$$\text{Re}[\lambda(A_{clr})] < 0, \quad (4.4)$$

where

$$A_{clr} =$$

$$\begin{bmatrix} A_{22} - K_r A_{12} - K_r B_1 F_2 & A_{21} - K_r B_1 F_1 & K_r A_{12} \\ -B_1 F_2 & A_{11} - B_1 F_1 & A_{12} \\ -B_2 F_2 & A_{21} - B_2 F_1 & A_{22} \end{bmatrix} \quad (4.5)$$

2. *ELTRI or ALTRI.* The achieved loop

transfer function $L_{cr}(j\omega)$,

$$L_{cr}(j\omega) = C_{cr}(j\omega)P(j\omega). \quad (4.6)$$

is either exactly or approximately equal to $L(j\omega)$, where $C_{cr}(s)$ denotes the transfer function of the compensator (i.e. the transfer function from y to $-\hat{u}$).

3. *Open-loop stability of the compensator.* The compensator is open-loop asymptotically stable, i.e.

$$\operatorname{Re}[\lambda(A_r)] < 0. \quad (4.7)$$

Besides satisfying the above technical conditions, as in the previous section, one expects that the value of gain needed for a certain degree of LTR is much smaller than that required by the reduced order observer-based controller. We have the following lemma:

Lemma 10. $E_{cr}(s)$, the error between the target loop transfer function $L(s)$ and that realized by the reduced order compensator, is given by

$$E_{cr}(s) = L(s) - L_{cr}(s) = M_r(s), \quad (4.8)$$

where

$$M_r(s) = F_r(\Phi_{22}^{-1} + K_r A_{12})^{-1} B_r. \quad (4.9)$$

Proof. See Appendix F.

Remark 11. The expression for $E_{cr}(s)$ is identical to the corresponding one for the full order compensator, see (3.8), except that now $M_r(s)$ takes the place of $M(s)$. Also $M_r(s)$ is the same as defined in (2.20) for the case of reduced order observer-based controller. We also note that the two expressions for the error between the required and the achieved loop transfer functions, one for the conventional reduced order observer design (2.19) and the other for the new compensator design (4.8), again differ significantly. Thus as we expect from Theorems 7 and 8, this leads to an overwhelming advantage in favor of the new reduced order compensator in contrast to a reduced order observer-based controller.

We have the following two theorems.

Theorem 9. Consider the closed-loop system as depicted in Fig. 5. Then both the asymptotic stability of the closed-loop system and ELTRI can be achieved under the following conditions:

1. The given plant has all its infinite zeros of order one.
2. The given plant is left invertible and is of minimum phase.

Moreover, under the above conditions, K_r can

be selected as an element of \mathcal{K}_{cr} . Also, the eigenvalues of A_{cl} are given by

$$\lambda(A_{22} - K_r A_{12}) \text{ and } \lambda(A - BF).$$

Furthermore, the developed compensator is always open-loop asymptotically stable.

Proof. Under the given conditions, K_r can be selected as an element of \mathcal{K}_{cr} and hence $M_r(s) \approx 0$. Thus ELTRI is achieved. Also, it is evident that the compensator is open-loop asymptotically stable. The closed-loop stability of Fig. 5 is given in Appendix G.

Theorem 10. Consider the closed-loop system as depicted in Fig. 5. Assume that the given plant is left invertible and is of minimum phase. Select the gain K_r which is parameterized in terms of a tuning parameter σ , as an element of $\mathcal{K}_{cr}(\sigma)$. Then ALTRI is achieved as $\sigma \rightarrow \infty$. Furthermore there exists a σ_2 such that the closed-loop system is asymptotically stable for all $\sigma > \sigma_2$. More specifically, as $\sigma \rightarrow \infty$, eigenvalues of A_{cl} are given by

$$\lambda(A_{22} - K_r(\sigma)A_{12}) + O(1/\sigma)$$

and

$$\lambda(A - BF) + O(1/\sigma).$$

Also, the developed compensator is always open-loop asymptotically stable.

Proof. The results of achieving ALTRI and open-loop asymptotic stability of the compensator are obvious. The proof of closed-loop stability of Fig. 5 is given in Appendix H.

As is clear by now, one often seeks an ALTRI design which as we know is asymptotic where the degree of recovery depends on the size of gain. Both the conventional reduced order observer-based controller and our new reduced order compensator are capable of achieving ALTRI. As expected, the following theorem, however, shows that for the same value of gain, the new compensator achieves a much better degree of recovery than the observer-based controller.

Theorem 11. Let $K_r(\sigma)$ be an element of $\mathcal{K}_{cr}(\sigma)$. Assume also that the same gain $K_r(\sigma)$ is used for both the reduced order observer-based controller and the reduced order compensator. Let σ be such that $\sigma_{\max}[M_r(j\omega)]$ is small (say, $\ll 1$) for all ω . Furthermore, assume that (3.15) is true. Then for all $\omega \in D_r$, the mismatch between the target loop transfer function and the one achieved by the reduced order compensator is always less than the corresponding one achieved by the reduced order observer-based controller.

More specifically, we have

$$\sigma_{\max}[E_{cr}(j\omega)] \gg \sigma_{\max}[E_{cr}(j\omega)] \quad \text{for all } \omega \in D_c, \quad (4.10)$$

where $E_{cr}(s)$ is as in (4.8) and $E_{cr}(s)$ is as in (2.19).

Proof. The proof follows along the same lines as that of Theorem 7.

As in the previous section, we now turn our attention to developing bounds on sensitivity and complementary sensitivity functions. Let $S_r(s)$ and $T_r(s)$ be the generated sensitivity and complementary sensitivity functions in the configuration of Fig. 5 when the loop is broken at the input point of the plant,

$$S_r(s) = [I_m + C_{cr}(s)P(s)]^{-1}$$

and

$$\begin{aligned} T_r(s) \\ = I_m - S_r(s) = [I_m + C_{cr}(s)P(s)]^{-1} C_{cr}(s)P(s). \end{aligned}$$

We have the following result analogous to Theorem 8.

Theorem 12. Consider the configuration of Fig. 5. Assume that (3.15) is true. We have the following bounds on all singular values $i = 1$ to m of $S_r(j\omega)$ and $T_r(j\omega)$:

$$\begin{aligned} \frac{|\sigma_i[S_r(j\omega)] - \sigma_i[S_r(j\omega)]|}{\sigma_{\max}[S_r(j\omega)]} \\ \leq \frac{\sigma_{\max}[M_r(j\omega)]}{\sigma_{\min}[F\Phi(j\omega)B] - \sigma_{\max}[M_r(j\omega)] - 1} \\ \ll \sigma_{\max}[M_r(j\omega)] \quad \text{for all } \omega \in D_c, \end{aligned} \quad (4.11)$$

and

$$\begin{aligned} \frac{|\sigma_i[T_r(j\omega)] - \sigma_i[T_r(j\omega)]|}{\sigma_{\max}[S_r(j\omega)]} \\ \leq \frac{\sigma_{\max}[M_r(j\omega)]}{\sigma_{\min}[F\Phi(j\omega)B] - \sigma_{\max}[M_r(j\omega)] - 1} \\ \ll \sigma_{\max}[M_r(j\omega)] \quad \text{for all } \omega \in D_c. \end{aligned} \quad (4.12)$$

Proof. It follows along the same lines as that of Theorem 8.

Remark 12. Remarks similar to 7–10 are obviously true even for reduced order compensators.

5. FULL AND REDUCED ORDER COMPENSATORS—ELTRO AND ALTRO

The results for the case when the target open-loop transfer functions are specified at the output point of the plant can be obtained by

dualizing those for the case when the target open-loop transfer functions are specified at the input point of the plant. However, one has to interpret duality in a proper manner and this was discussed earlier in Section 2.3. All this discussion also applies to compensator design. All one has to do is to use either full or reduced order compensator design of Sections 3 and 4 to achieve LTR in the third step of the design algorithm discussed in Section 2.3. The remaining steps of the design algorithm given in Section 2.3 remain intact.

6. EXAMPLES

Examples are presented in this section comparing the new compensator with the conventional observer approach. These examples are worked out using the software reported by Chen *et al.* (1989). Clearly all the examples support the theoretical development given earlier and demonstrate that the new compensator approach is much better than the conventional observer approach in all cases, namely, (a) when the performance specifications are reflected either at the input or at the output point of the plant, and (b) whether the full or reduced order compensator is used.

Most often in the literature, the maximum and minimum singular value graphs of the target and achieved loop transfer matrices are drawn with respect to ω and are then compared. These graphs could be misleading. Although the singular values of target and achieved loop transfer matrices may match perfectly, the difference or mismatch between them could be very high owing to the phase difference between them. This has been pointed out by Ridgely and Banda (1986) in an example. The best way is to check the singular values of the *mismatch function* between the target and achieved loop transfer matrices.

To show the effects of the phase difference, consider the example of Ridgely and Banda (1986)

$$\dot{x} = Ax + Bu + \Gamma \zeta$$

$$\begin{bmatrix} 0 & 1 \\ 4 & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 35 \\ -61 \end{bmatrix} \zeta \quad (6.1)$$

and

$$y = Cx + \eta = [2 \quad 1]x + \eta. \quad (6.2)$$

The state feedback law is selected as

$$u = -Fx = -[50 \quad 10]x. \quad (6.3)$$

The full order observer gain $K(\sigma)$ is obtained by solving the filter Riccati equation,

$$A\Sigma + \Sigma A' + Q(\sigma) - \Sigma C' C \Sigma = 0 \quad (6.4)$$

and

$$K(\sigma) = \Sigma C'$$

where

$$Q(\sigma) = \Gamma \Gamma' + \sigma^2 B B'$$

The magnitude plots of the target and achieved loop transfer function when σ^2 takes values 0; 500; 2500; 3600; 8100; and 250,000 are presented in Fig. 6(a). As σ^2 begins to increase, the low frequency region of the achieved loop transfer function begins to approach the target loop, while the high frequency region remains virtually unchanged. As σ^2 takes the value 3000, the low frequency region also almost matches the target loop. At $\sigma^2 = 3600$ as shown in Fig. 6(a), the target and achieved loop transfer function magnitudes are almost "tight" together. However, as shown in Fig. 6(b), the phases are about 180° apart in the low frequency region. This shows that no recovery has been achieved

TABLE 1(a). SUPREMUM OF MAXIMUM SINGULAR VALUES OF MISMATCH FUNCTIONS OVER FREQUENCIES PLOTTED

	Tuning parameter	Supremum $\sigma_{\max}(E_0(j\omega))$	Supremum $\sigma_{\max}(E_c(j\omega))$
Case 1	$\sigma^2 = 500$	20.1627	8.0747
Case 2	$\sigma^2 = 10^3$	21.6324	5.4534
Case 3	$\sigma^2 = 10^4$	55.7910	0.7910

TABLE 1(b). COMPARISON OF FULL ORDER OBSERVER BASED CONTROLLER VS FULL ORDER COMPENSATOR FOR THE SAME DEGREE OF RECOVERY

Degree of recovery $\sup \sigma_{\max}[E_0(j\omega)] = \sup \sigma_{\max}[E_c(j\omega)] = 0.7910$ for $10^{-2} \leq \omega \leq \infty$ rad/s		
	Observer-based controller	Full order compensator
Gain norm	353.4295	84.8997
Eigenvalues	-1.8664 -370.9527	-2.0603 -100.4328
Bandwidth	5170 rad/s	1682 rad/s

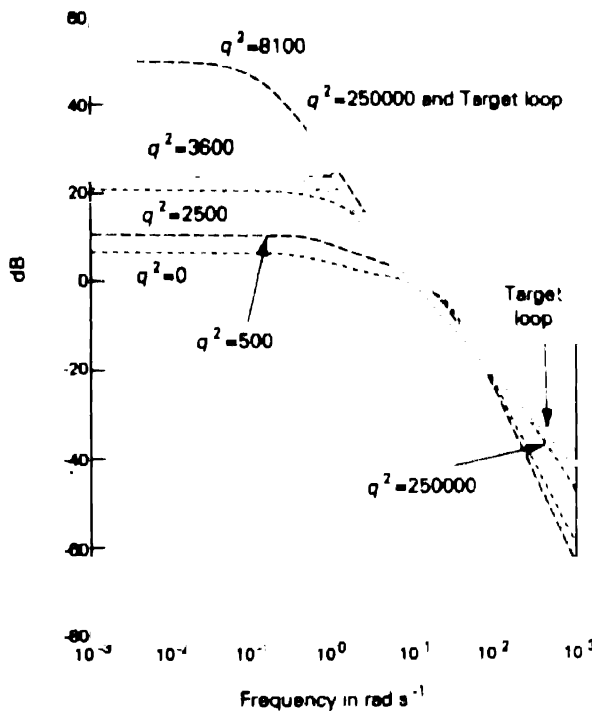


FIG. 6(a). Singular values of target loop and design loops

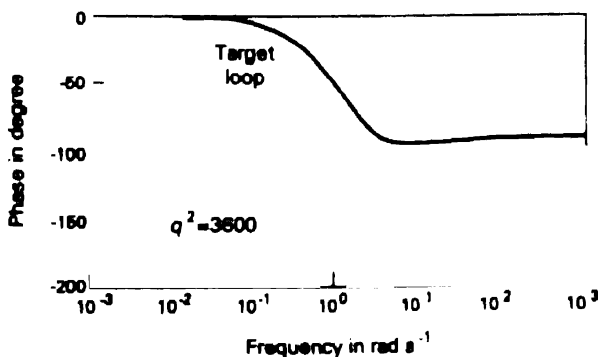


FIG. 6(b). Phase responses of target loop and design loop with $q^2 = 3600$.

at $\sigma^2 = 3600$. It takes a σ^2 of 250,000 to achieve the needed recovery.

In what follows, for each example, we present the traditional maximum and minimum singular value graphs of the target and achieved loop transfer matrices. However, in view of the above discussion, the maximum singular value graphs of the mismatch function are also separately given. Also, a tabular column presents the supremum of the maximum singular value of the mismatch function with respect to ω over the frequency range of interest. All of the above data relate to the comparison between the observer-based controller and our new compensator when both of them use the same value of gain. Another method of comparison is to give the value of gain, eigenvalues and bandwidth of both the observer-based controller and the compensator in order that both of them achieve the same degree of recovery as measured by the supremum of the maximum singular value of the correspondingly generated mismatch function. Another tabular column shows this information. Also, for a chosen supremum of maximum singular value, a graph shows the variation of maximum singular value of the observer-based controller and that of the compensator with respect to ω over the frequency range of interest. From all these data, it is easy to see that the new compensator approach has better recovery properties than the conventional observer approach.

Example 1 (Full order ALTRI). Consider the example in Doyle and Stein (1979) [and see Table 1 (a, b) and Fig. 7 (a, b)].

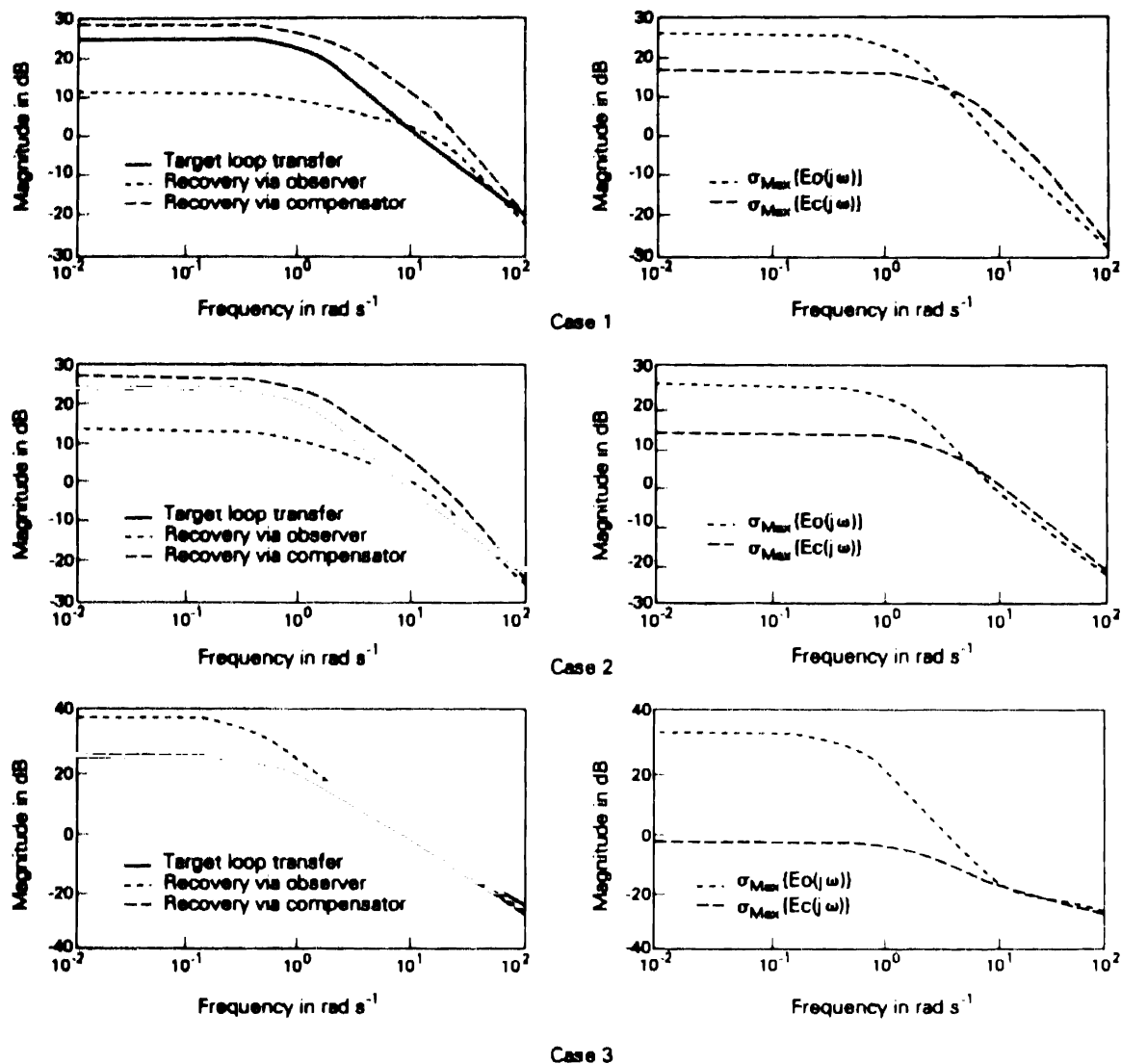


FIG. 7(a). Frequency responses for all the cases given in Table 1(a)

Plant:

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -3 & -4 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u,$$
$$y = \begin{bmatrix} 2 & -1 \end{bmatrix} x.$$

State feedback gain:

$$F = \begin{bmatrix} 50 & -10 \end{bmatrix}.$$

Example 2 (Full order ALTRO). Consider the following example in Ridgely and Banda (1986) [and see Table 2 (a, b) and Fig. 8 (a, b)]

$$A = \begin{bmatrix} -0.08527 & -0.0001423 & -0.9994 & 0.04142 & 0 & 0.1862 \\ -46.86 & -2.757 & 0.3896 & 0 & -124.3 & 128.6 \\ -0.4248 & -0.06224 & -0.06714 & 0 & -8.792 & -20.46 \\ 0 & 1 & 0.0523 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -20 & 0 \\ 0 & 0 & 0 & 0 & 0 & -20 \end{bmatrix}$$

TABLE 2(a). SUPRENUM OF MAXIMUM SINGULAR VALUES OF MISMATCH FUNCTIONS OVER FREQUENCIES PLOTTED

Supremum $\sigma_{\max}(E_o(j\omega))$	Supremum $\sigma_{\max}(E_c(j\omega))$
471.9951	0.2398

TABLE 2(b). COMPARISON OF FULL ORDER OBSERVER-BASED CONTROLLER VS FULL ORDER COMPENSATOR FOR THE SAME DEGREE OF RECOVERY

	Degree of recovery	
$\sup \sigma_{\max}[E_0(j\omega)] \approx \sup \sigma_{\max}[E_c(j\omega)] \approx 0.2398$	for $10^{-3} \leq \omega < \infty \text{ rad s}^{-1}$	
	Observer-based controller	Full order compensator
Gain norm	3.1623×10^9	317.4681
Eigenvalues	-158.15 -19923 $-76736 \pm j76737$ $-9956 \pm j17244$	-104.69 -51.17 $-52.5 \pm j89.62$ $-26.8 \pm j42.71$
Bandwidth	$4.78 \times 10^{10} \text{ rad/s}^{-1}$	5288 rad/s ⁻¹

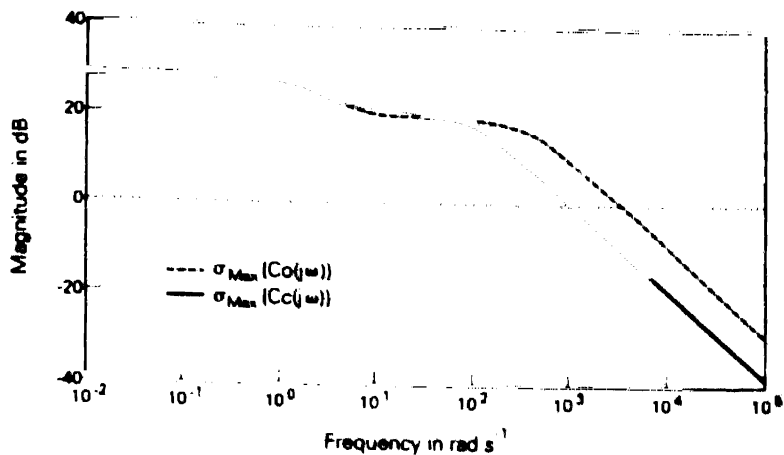


FIG. 7(b) Maximum singular values of full order observer based controller and compensator given in Table 1(b)

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 0 & 0 & 20 \end{bmatrix}^T$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

For this example, the observer gain is given, instead of the state feedback gain, to meet all the design specifications for the loop broken at the output of plant.

$$K = \begin{bmatrix} 4.20 & -18.17 & -9.92 & -1.19 & 0.0181 & 0.1149 \\ -1.19 & 55.81 & -0.60 & 10.49 & 0.3330 & 0.3306 \end{bmatrix}^T$$

Example 3 (reduced order ALTRI). Consider the example in Sogaard-Andersen (1987b) [and see Table 3 and Fig. 9]

$$A = \begin{bmatrix} 2 & 0 & 1 & 0 & 0 & 1 & 1 \\ -1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 3 & 2 & -1 & 0 & 0 & -2 & 1 \\ 2 & -2 & 0 & -4 & 2 & 0 & -1 \\ 0 & 2 & 3 & 0 & -2 & 1 & -1 \\ 1 & 0 & 2 & -3 & 2 & 2 & 0 \\ -1 & -1 & 1 & 0 & 0 & -1 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}$$

$$C = [I_3 \quad 0_{3 \times 4}]$$

The state feedback gain is an *LQ*-design with weights $Q = I_7$ and $R = 10^{-3}I_3$.

TABLE 3 SUPREMUM OF MAXIMUM SINGULAR VALUES OF MISMATCH FUNCTIONS OVER FREQUENCIES PLOTTED

	Tuning parameter	Supremum $\sigma_{\max}(E_o(j\omega))$	Supremum $\sigma_{\max}(E_c(j\omega))$
Case 1	$\alpha = 10$	203.0387	19.9622
Case 2	$\alpha = 100$	136.1517	2.1660

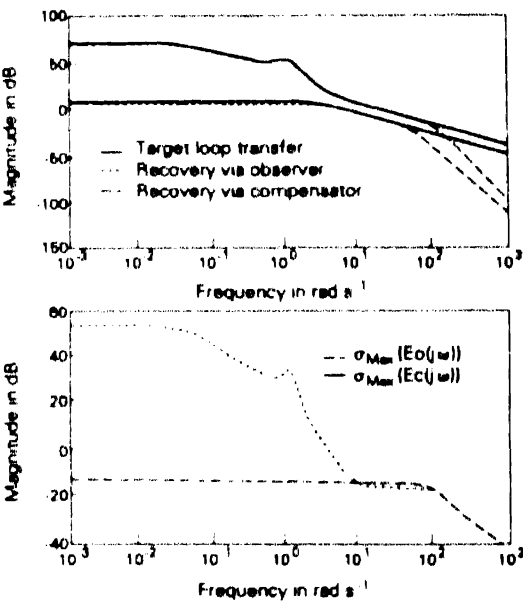


FIG. 8(a) Frequency responses for the case given in Table 2(a)

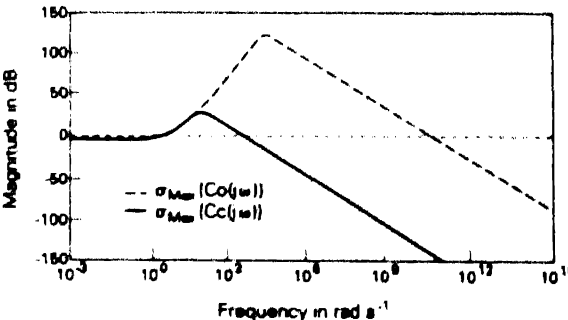


FIG. 8(b) Maximum singular values of full order observer based controller and compensator given in Table 2(b)

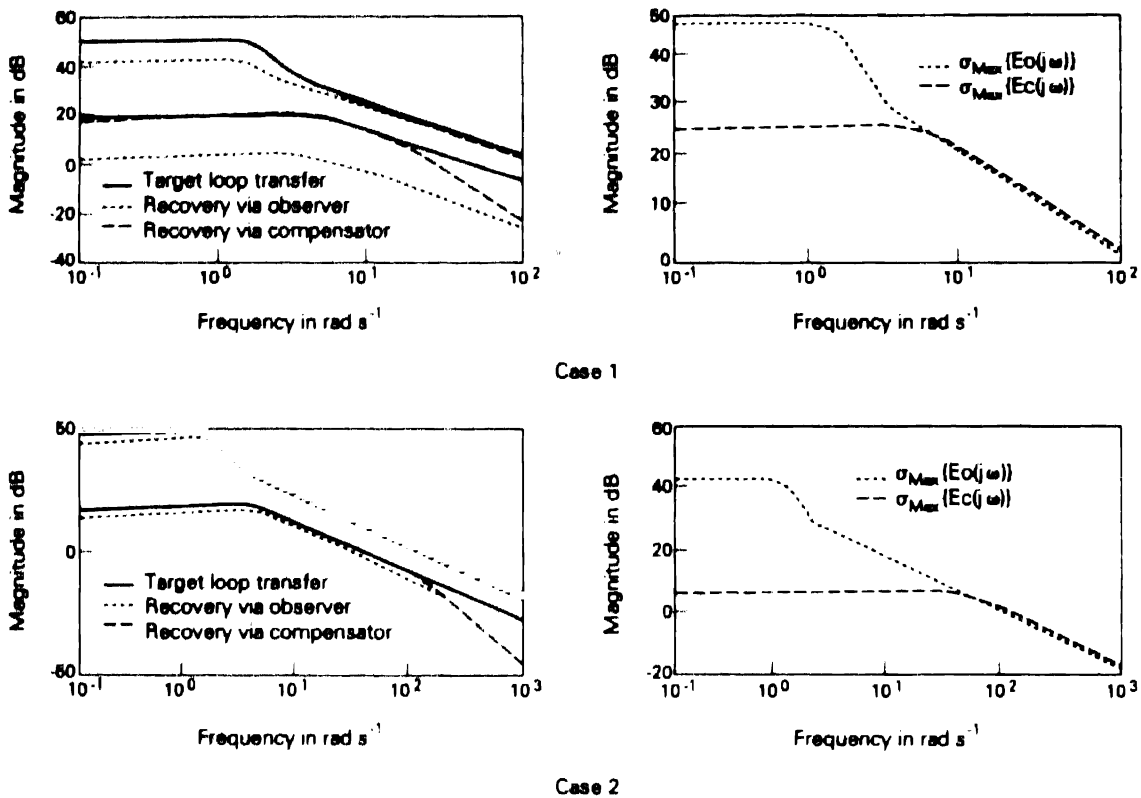


FIG. 9 Frequency responses for both the case given in Table 3(a)

7 CONCLUSIONS

The loop transfer recovery (LTR) methods using observer or Kalman filter-based controllers are streamlined and the theory of both full and reduced order observers is brought to the same framework. It is shown that either exact or approximate LTR can be accomplished iff $M(j\omega)$ [or equivalently $M_r(j\omega)$ for the reduced order observers] either exactly or approximately zero for all ω . The term $M(s)$ [or $M_r(s)$] has a physical interpretation; it is the transfer function from the point where the input u of the plant is fed to the observer based controller to the output point $-\hat{u}$ of the controller. Also, the conditions for ELTR are presented directly in terms of the given system matrices A , B and C and the state feedback gain F . The methods of calculating the needed observer gain for both full and reduced order observer based controllers to achieve LTR are presented. The singular value bounds on the difference between the achieved and the target sensitivity and complementary sensitivity functions are developed when both full and reduced order observer based controllers are used. The duality between the two cases when the design specifications reflect at the input or at the output point of the plant is discussed. One has to interpret this duality carefully. From the view point of design methodology, the two

cases are completely dual and this duality holds for either full or reduced order observer-based controllers or in fact for any other controllers or compensators.

A new compensator structure for loop transfer recovery either at the input or at the output point of the plant is proposed. It could be either full or reduced order type. The compensator is structurally different from the observer in the sense that no link from the input point of the plant to the controller is used. This omission of the link from the input point of the plant to the controller has a profound effect on all aspects of the loop transfer recovery. It results in an open-loop stable compensator. Also, the closed-loop stability can be guaranteed. More importantly, the value of gain required for a given degree of LTR is orders of magnitude less than what is required in the conventional approach. Also, singular value bounds on sensitivity and complementary sensitivity functions illustrate that the proposed compensator has better recovery properties than the conventional observer based controller. These advantages reflect in various ways. First, the woes of saturation are either eliminated or at least dampened. The controller band-width is reduced and consequently the control signal to noise ratio at the input point of the plant is increased. All

these claims are theoretically obvious from our development. Also, numerical examples illustrate the same.

A fundamental assumption throughout this paper has been that the given plant is of minimum phase. We are presently in the process of developing compensators for nonminimum phase plants where obviously, the structures given in this paper will not work out since in general nonminimum phase plants might not be stabilizable by stable compensators. Hence we are looking at some other appropriate structures to deal with nonminimum phase systems. These results will be reported in a forthcoming paper.

Acknowledgement—The work of B. M. Chen and A. Saberi is supported in part by National Science Foundation under grant no. ECS 8618953 and in part by Boeing Commercial Airplanes.

REFERENCES

- Athans, M. (1986). A tutorial on LQG/LTR methods. *Proc. CDC*, Athens, Greece.
- Baumgartner, C. E., H. D. Geering, C. H. Onder and F. Shafai (1986). Robust multivariable idle speed control. *Proc. ACC*, Seattle, Washington, pp. 258–265.
- Chen, B. M., P. Bingulac and A. Saberi (1989). A C/A-D package for LTR design via time-scale structure and eigenvalue assignment. *Proc. 27th Annual Allerton Conf. on Communications, Control and Computing*, 439–440.
- Dowdle, J. R., G. Stein and N. R. Sandell (1982). Minimal order robust observer-based compensator. *Proc. CDC*, Orlando, Florida, pp. 894–896.
- Doyle, J. C. and G. Stein (1979). Robustness with observers. *IEEE Trans. Aut. Control*, **AC-24**, 607–611.
- Doyle, J. C. and G. Stein (1981). Multivariable feedback design: concepts for a classical/modern synthesis. *IEEE Trans. Aut. Control*, **AC-26**, 4–16.
- Friedland, B. (1986). *Control System Design—An Introduction to State Space Methods*. McGraw-Hill, New York.
- Goodman, G. C. (1984). *The LQG/LTR method and discrete-time control systems*. Report No. LIDS-TN-1392, MIT, MA.
- Kazerooni, H. and P. K. Houpt (1986). On the loop transfer recovery. *Int. J. Control*, **43**, 981–986.
- Khalil, H. K. (1981). On the robustness of output feedback control methods to modelling errors. *IEEE Trans. Aut. Control*, **AC-26**, 524–526.
- Khalil, H. K. (1984). A further note on the robustness of output feedback control methods to modelling errors. *IEEE Trans. Aut. Control*, **AC-29**, 861–862.
- Kwakernaak, H. (1969). Optimal low sensitivity linear feedback systems. *Automatica*, **5**, 279–285.
- Kwakernaak, H. and R. Sivan (1972). *Linear Optimal Control Systems*, Wiley, Chichester.
- Madiwale, A. N. and D. E. Williams (1985). Some extensions of loop transfer recovery. *Proc. 1985 ACC*, Boston, MA, pp. 790–795.
- Matson, C. L. and P. S. Maybeck (1987). On an assumed convergence result in the LQG/LTR technique. *Proc. 26th CDC*, Los Angeles, CA, pp. 951–952.
- O'Reilly, J. (1983). *Observers for Linear Systems*. Academic Press, London.
- Ridgely, D. B., and S. S. Banda (1986). *Introduction to robust multivariable control*. Report No. AFWA1-TR-85-3102, Flight Dynamics Laboratories, Wright-Patterson Air Force Base, OH.
- Saberi, A. and P. Sannuti (1988). Observer design for loop transfer recovery and for uncertain dynamical systems. *Proc. ACC*, Atlanta, Georgia, pp. 803–808. Also (1990) *IEEE Trans. Aut. Control*, **AC-35**, 878–897.
- Sannuti, P. and A. Saberi (1987). A special coordinate basis of multivariable linear systems—finite and infinite zero structure, squaring down and decoupling. *Int. J. Control*, **45**, 1655–1704.
- Shaked, U. and N. Karcanias (1976). The use of zeros and zero-directions in model reduction. *Int. J. Control*, **23**, 113–135.
- Shaw, L. (1971). Pole placement, stability and sensitivity of dynamic compensators. *IEEE Trans. Aut. Control*, **AC-16**, 210.
- Sogaard-Andersen, P. (1987a). Comments on 'On the loop transfer recovery'. *Int. J. Control*, **45**, 369–374.
- Sogaard-Andersen, P. (1987b). Loop transfer recovery with minimal order observers. *Proc. 26th CDC*, Los Angeles, CA, pp. 933–938.
- Sogaard-Andersen, P. (1987c). Explicit solution to the problem of exact loop transfer recovery. *Proc. 1987 ACC*, Minneapolis, MN, pp. 150–151.
- Sogaard-Andersen, P. and H. H. Niemann (1989). Trade-offs in LTR-based feedback design. *Proc. 1989 ACC*, Pittsburgh, PA, pp. 922–928.
- Stein, G. and M. Athans (1987). The LQG/LTR procedure for multivariable feedback control design. *IEEE Trans. Aut. Control*, **AC-32**, 105–114.
- Vidyasagar, M. (1985a). Robust stabilization of singularly perturbed systems. *Syst. Control Lett.*, **5**, 413–418.
- Vidyasagar, M. (1985b). *Control System Synthesis: A Factorization Approach*. Springer, New York.

APPENDIX A PROOF OF THEOREM 1 AND CALCULATION OF GAIN K TO ACHIEVE ELTRI

Under the conditions (2) and (3) of Theorem 1, a theorem of Sannuti and Saberi (1987) implies that there exist nonsingular transformations T_1 , T_2 and T_3 such that

$$\begin{aligned} (T_1^{-1}A_1T_1 - T_1^{-1}E_1T_1)^{-1}T_1^{-1}B_1 &= T_1^{-1}A_1 \\ T_1^{-1}C_1 &= [C_{1a}^T \ C_{1b}^T \ C_1^T]^T \\ T_1^{-1}A_1T_1 &= A_{1a}T_a + T_{ab}T_b + T_{at}T_t \\ T_1^{-1}B_1 &= A_{1b}T_b + T_{bt}T_t, \quad T_{ab} = C_{1a}T_a \\ T_1^{-1}C_1 &= T_{at}T_a + T_{bt}T_b + T_{tt}T_t, \quad T_{tt} = C_{1t} \end{aligned} \quad (A.1)$$

Here, the pair (A_{1b}, C_{1t}) is observable. Furthermore, $\lambda(A_{1a})$ are the invariant zeros of the given plant and hence in view of condition (2) of Theorem 1, they are in the left half s plane.

Since (A_{1b}, C_{1t}) is observable, one can select a gain K_{1b} such that $\lambda(A_{1b} - K_{1b}C_{1t})$ are in the desired locations in \mathbb{C} . Also, one can always choose a gain K_{1t} such that $\lambda(E_{1t} - K_{1t})$ are in the desired locations in \mathbb{C} . Now choose a gain K as,

$$\begin{aligned} K_{1a} &= T_{at} \\ K_{1b} &= T_{ab} \\ K_{1t} &= K_{1t} \end{aligned} \quad (A.2)$$

where K_{1b} is an arbitrary matrix with appropriate dimensions. Finally let

$$K = T_1 K_1 T_1^{-1}$$

All such gains K with K_{1b} arbitrary form the class of gains \mathcal{K}_L . Due to the special structure of matrices in (A.1) and in view of (A.2), it is straightforward to verify that $A - KC$ has eigenvalues in \mathbb{C}^- and that

$$F(\Phi^{-1} + KC)^{-1}B = 0$$

whenever $FB = 0$. Hence in view of Lemmas 1 and 2, ELTRI is achieved.

APPENDIX B: PROOF OF LEMMA 6

In the reduced order observer-based feedback control system of Fig. 4, at first we want to evaluate the loop transfer function $L_m(s)$ when the loop is broken at the input point of the plant. For this purpose, consider the plant input u and the controller output \hat{u} as two separate variables. Then from (2.15) to (2.17),

$$\dot{\hat{x}}_2 = (A_{22} - K_r A_{12})\hat{x}_2 + A_{21}x_1 + B_r\hat{u} + K_r(x_1 - A_{11}x_1).$$

Hence

$$\begin{aligned}\hat{x}_2(s) &= (\Phi_{22}^{-1} + K_r A_{12})^{-1} [(A_{21} + K_r \Phi_{11}^{-1})x_1(s) + B_r \hat{u}(s)] \\ &\quad - \hat{u}(s) = F_1 x_1(s) + F_2 (\Phi_{22}^{-1} + K_r A_{12})^{-1} \\ &\quad \times [(A_{21} + K_r \Phi_{11}^{-1})x_1(s) + B_r \hat{u}(s)]\end{aligned}$$

Thus

$$\begin{aligned}[I_m + F_2 (\Phi_{22}^{-1} + K_r A_{12})^{-1} B_r] [F_1 x_1(s) + F_2 \hat{x}_2(s)] \\ \approx F_1 x_1(s) + F_2 (\Phi_{22}^{-1} + K_r A_{12})^{-1} (A_{21} + K_r \Phi_{11}^{-1}) x_1(s),\end{aligned}$$

and therefore

$$L_m(s) = [I_m + M_r(s)]^{-1} [F_1 + F_2 (\Phi_{22}^{-1} + K_r A_{12})^{-1} (K_r \Phi_{11}^{-1} + A_{21})] P(s), \quad (B.1)$$

where

$$M_r(s) = F_2 (\Phi_{22}^{-1} + K_r A_{12})^{-1} B_r.$$

We will next simplify some expressions. Using (2.12),

$$\begin{aligned}(K_r \Phi_{11}^{-1} + A_{21}) P(s) \\ \approx (K_r \Phi_{11}^{-1} + A_{21}) (\Phi_{11} B_1 + \Phi_{11} A_{12} H_2(s)) \\ \approx K_r B_1 + B_2 + A_{21} \Phi_{11} B_1 + B_2 + (K_r + A_{21} \Phi_{11}) A_{12} H_2(s)\end{aligned} \quad (B.2)$$

But from (2.14)

$$A_{21} \Phi_{11} B_1 + B_2 = (\Phi_{22}^{-1} - A_{21} \Phi_{11} A_{12}) H_2(s). \quad (B.3)$$

Thus (B.2) and (B.3) imply that

$$(K_r \Phi_{11}^{-1} + A_{21}) P(s) \approx -B_r + (\Phi_{22}^{-1} + K_r A_{12}) H_2(s) \quad (B.4)$$

Using (B.4) and (2.13),

$$\begin{aligned}F_2 (\Phi_{22}^{-1} + K_r A_{12})^{-1} (K_r \Phi_{11}^{-1} + A_{21}) P(s) \\ \approx -F_2 (\Phi_{22}^{-1} + K_r A_{12})^{-1} B_r + F_2 H_2(s) \\ \approx L(s) - M_r(s)\end{aligned} \quad (B.5)$$

Now in view of (B.1) and (B.5),

$$L_m(s) = [I_m + M_r(s)]^{-1} [L(s) - M_r(s)]$$

Thus we have

$$\begin{aligned}E_m(s) = L(s) - L_m(s) \\ \approx [I_m + M_r(s)]^{-1} [(I_m + M_r(s))L(s) - L(s) + M_r(s)] \\ \approx [I_m + M_r(s)]^{-1} M_r(s)(L_m + L(s)) \\ \approx M_r(s)[I_m + M_r(s)]^{-1} (L_m + L(s)).\end{aligned}$$

APPENDIX C: PROOF OF THEOREM 3

Without loss of generality, we will assume that the given plant is in the form of a special coordinate basis as in (A.1) (see Appendix A). Then partitioning the state variable x_p as $x_p = [x_{p1}^T, x_{p2}^T]^T$ with $v_i = x_{p1i}$, we can write the matrices A , B and C characterizing (A.1) as

$$\begin{aligned}A = \begin{bmatrix} E_f & E_{k1} & E_{k2} & E_a \\ L_{k1} & A_{k11} & A_{k12} & 0 \\ L_{k2} & A_{k21} & A_{k22} & 0 \\ L_{a1} & L_{a2} & 0 & A_{aa} \end{bmatrix} \\ C = \begin{bmatrix} I_m & 0 & 0 \\ 0 & I_p & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (C.1)\end{aligned}$$

The triple (A, B, C) in (C.1) assumes that the condition 1 of Theorem 3 is true. Then in view of (C.1), (2.9) and (2.10), we have

$$\begin{aligned}A_{22} = \begin{bmatrix} A_{k22} & 0 \\ 0 & A_{aa} \end{bmatrix}, \quad A_{12} = \begin{bmatrix} E_{k2} & E_a \\ A_{k12} & 0 \end{bmatrix}, \\ B_1 = \begin{bmatrix} I_m \\ 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.\end{aligned}$$

Also, it can be easily seen that the pair (A_{k22}, A_{k12}) is observable (Saberi and Sannuti, 1988). Hence there exists a K_{k2} such that $\lambda(A'_{k22})$ are in the desired locations in \mathbb{C} where $A'_{k22} = A_{k22} - K_{k2} A_{k12}$. Now consider a reduced order observer gain matrix K_r as

$$K_r = \begin{bmatrix} 0 & K_{k2} \\ 0 & K_{aa} \end{bmatrix} \quad (C.2)$$

where K_{aa} is arbitrary. It is then simple to verify that A_r ,

$$A_r = A_{22} - K_r A_{12} = \begin{bmatrix} A'_{k22} & 0 \\ -K_{aa} A_{k12} & A_{aa} \end{bmatrix}$$

is a stable matrix provided that the given plant is of minimum phase, i.e. $\text{Re } \lambda(A_{aa}) < 0$. Furthermore, we note that

$$B_r = B_2 - K_r B_1 = 0$$

This in view of (2.22) shows that ELTRI is achieved. Also, we note that all gains K_r as in (C.2) with K_{aa} arbitrary form the class of gains K_{er} .

APPENDIX D: PROOF OF LEMMA 8

From (2.19), we have

$$\begin{aligned}E_m(s) = F\Phi B + C_m(s)P(s) \\ \approx M_r(s)[I_m + M_r(s)]^{-1} (I_m + F\Phi B),\end{aligned}$$

and hence

$$\begin{aligned}I_m + C_m(s)P(s) \\ \approx I_m + F\Phi B - E_m(s) \\ \approx I_m + F\Phi B - M_r(s)[I_m + M_r(s)]^{-1} (I_m + F\Phi B) \\ \approx [I_m + M_r(s)]^{-1} (I_m + F\Phi B)\end{aligned}$$

Thus

$$S_r(s) = S_f(s)[I_m + M_r(s)] \quad (D.1)$$

Then using singular value inequalities, we have for each $i = 1$ to m ,

$$\sigma_i[S_{ir}(j\omega)] \leq \sigma_i[S_f(j\omega)] + \sigma_{\max}[S_f(j\omega)M_r(j\omega)],$$

and thus

$$\sigma_i[S_{ir}(j\omega)] - \sigma_i[S_f(j\omega)] \leq \sigma_{\max}[S_f(j\omega)] \sigma_{\max}[M_r(j\omega)]. \quad (D.2)$$

Now rewriting (D.1) as,

$$S_f(s) \approx S_{ir}(s) - S_f(s)M_r(s),$$

we have for each $i = 1$ to m ,

$$\sigma_i[S_f(j\omega)] - \sigma_i[S_{ir}(j\omega)] \leq \sigma_{\max}[S_f(j\omega)] \sigma_{\max}[M_r(j\omega)]. \quad (D.3)$$

Then in view of (D.2) and (D.3), we get

$$|\sigma_i[S_{ir}(j\omega)] - \sigma_i[S_f(j\omega)]| \leq \sigma_{\max}[M_r(j\omega)].$$

Next in view of (D.1),

$$T_m(s) = I_m - S_m(s) = T_f(s) - S_f(s)M_r(s).$$

Now using singular value inequalities and proceeding as above, we get

$$|\sigma_i[T_m(j\omega)] - \sigma_i[T_f(j\omega)]| \leq \sigma_{\max}[M_r(j\omega)].$$

APPENDIX E: PROOF OF THEOREM 8

From (3.8), we have

$$E_r(s) = F\Phi B - C_r(s)P(s) = M(s),$$

and hence

$$\begin{aligned} I_m + C_r(s)P(s) &= I_m + F\Phi B - M(s) \\ &= [I_m - M(s)[I_m + F\Phi B]^{-1}][I_m + F\Phi B] \end{aligned}$$

Thus

$$\begin{aligned} S_r(s) &= S_r(s)[I_m - M(s)[I_m + F\Phi B]^{-1}]^{-1} \\ &= S_r(s)[I_m + M(s)[I_m + F\Phi B - M(s)]^{-1}] \\ &= S_r(s) + S_r(s)M(s)[I_m + F\Phi B - M(s)]^{-1}. \quad (\text{E.1}) \end{aligned}$$

Then using singular value inequalities, we have for each $i = 1$ to m ,

$$\begin{aligned} \sigma_i[S_r(j\omega)] &\leq \sigma_i[S_r(j\omega)] \\ &\quad + \sigma_{\max}\{S_r(j\omega)M(j\omega)[I_m + F\Phi(j\omega)B - M(j\omega)]^{-1}\}, \end{aligned}$$

or equivalently

$$\begin{aligned} \sigma_i[S_r(j\omega)] &\leq \sigma_i[S_r(j\omega)] \\ &\leq \sigma_{\max}\{S_r(j\omega)M(j\omega)[I_m + F\Phi(j\omega)B - M(j\omega)]^{-1}\} \quad (\text{E.2}) \end{aligned}$$

Now rewriting (E.1) as

$$S_r(s) = S_r(s) = S_r(s)M(s)[I_m + F\Phi B - M(s)]^{-1},$$

we have for each $i = 1$ to m ,

$$\begin{aligned} \sigma_i[S_r(j\omega)] &\leq \sigma_i[S_r(j\omega)] \\ &\quad + \sigma_{\max}\{S_r(j\omega)M(j\omega)[I_m + F\Phi(j\omega)B - M(j\omega)]^{-1}\}, \end{aligned}$$

or equivalently

$$\sigma_i[S_r(j\omega)] \leq \sigma_i[S_r(j\omega)] + \sigma_{\max}\{S_r(j\omega)M(j\omega)[I_m + F\Phi(j\omega)B - M(j\omega)]^{-1}\} \quad (\text{E.3})$$

Combining (E.2) and (E.3), we get for each $i = 1$ to m ,

$$\begin{aligned} |\sigma_i[S_r(j\omega)]| &\leq |\sigma_i[S_r(j\omega)]| \\ &\quad + \sigma_{\max}[S_r(j\omega)] \\ &\leq \sigma_{\max}[M(j\omega)]\sigma_{\max}[I_m + F\Phi(j\omega)B - M(j\omega)]^{-1} \\ &\leq \sigma_{\max}[M(j\omega)] \\ &\leq \sigma_{\min}[I_m + F\Phi(j\omega)B - M(j\omega)] \\ &\leq \sigma_{\max}[M(j\omega)] \\ &\leq \sigma_{\min}[F\Phi(j\omega)B] = \sigma_{\max}[M(j\omega)]^{-1} \\ &\leq \sigma_{\max}[M(j\omega)] \text{ for all } \omega \in D. \quad (\text{E.4}) \end{aligned}$$

The last step in (E.4) follows from (3.15). Next, in view of (E.1),

$$\begin{aligned} T_r(s) &= I_m - S_r(s) \\ &= I_m - S_r(s) = S_r(s)M(s)[I_m + F\Phi B - M(s)]^{-1} \\ &= T_r(s) = S_r(s)M(s)[I_m + F\Phi B - M(s)]^{-1} \end{aligned}$$

Now using singular value inequalities and proceeding as above, we get

$$\begin{aligned} |\sigma_i[T_r(j\omega)]| &\leq \sigma_i[T_r(j\omega)] \\ &\leq \sigma_{\max}[M(j\omega)] \text{ for all } \omega \in D, \end{aligned}$$

APPENDIX F: PROOF OF LEMMA 10

In the reduced order compensator-based feedback control system of Fig. 5, at first we want to evaluate the loop transfer function $L_{r,r}(s)$ when the loop is broken at the input point of the plant. For this purpose, consider the plant input u and the controller output \hat{u} as two separate variables. Then in view of (4.1) to (4.3),

$$\dot{w} = (A_{22} - K_r(\sigma)A_{12})\dot{w} + A_{21}\dot{x}_1 + K_r(\sigma)(\dot{x}_1 - A_{11}\dot{x}_1) \quad (\text{F.1})$$

Hence

$$w(s) = (\Phi_{22}^{-1} + K_r(\sigma)A_{12})^{-1}(A_{21} + K_r(\sigma)\Phi_{11}^{-1})P(s)u(s) \quad (\text{F.2})$$

Thus in view of (F.2),

$$\begin{aligned} -\hat{u}(s) &= F_1 x_1(s) + F_2 w(s) \\ &= [F_1 + F_2(\Phi_{22}^{-1} + K_r(\sigma)A_{12})^{-1}(A_{21} + K_r(\sigma)\Phi_{11}^{-1})]P(s)u(s) \quad (\text{F.3}) \end{aligned}$$

Now using (B.5),

$$L_{r,r}(s) = L(s) - M_r(s). \quad (\text{F.4})$$

Hence

$$L(s) - L_{r,r}(s) = M_r(s)$$

APPENDIX G: PROOF OF THEOREM 9

We first note the following

$$sI_n - A = \Phi^{-1} \begin{bmatrix} \Phi_{11}^{-1} & A_{12} \\ -A_{21} & \Phi_{22}^{-1} \end{bmatrix},$$

and hence

$$\Phi^{-1} \begin{bmatrix} 0 \\ I_{n-p} \end{bmatrix} = \Phi_a,$$

where

$$\Phi_a = \begin{bmatrix} -A_{12} \\ \Phi_{22}^{-1} \end{bmatrix}$$

Thus

$$F\Phi\Phi_a = F_2 \quad (\text{G.1})$$

Using A_{10} as in (4.5), we have the following series of reductions

$$\begin{aligned} \det[sI_{2n-p} - A_{10}] &= \det \begin{bmatrix} \Phi_{22}^{-1} + K_r A_{12} + K_r B_1 F_2 & -A_{21} + K_r B_1 F_1 & -K_r A_{12} \\ B_1 F_2 & \Phi_{11}^{-1} + B_1 F_1 & -A_{12} \\ B_2 F_2 & -A_{21} + B_2 F_1 & \Phi_{22}^{-1} \end{bmatrix} \\ &= \det \begin{bmatrix} \Phi_{22}^{-1} + K_r B_1 F_2 & -A_{21} + K_r B_1 F_1 & -K_r A_{12} \\ -A_{12} + B_1 F_2 & \Phi_{11}^{-1} + B_1 F_1 & -A_{12} \\ \Phi_{22}^{-1} + B_2 F_2 & -A_{21} + B_2 F_1 & \Phi_{22}^{-1} \end{bmatrix} \\ &= \det \begin{bmatrix} \Phi_{22}^{-1} + K_r B_1 F_2 & -A_{21} + K_r B_1 F_1 & -K_r A_{12} \\ -A_{12} + B_1 F_2 & \Phi_{11}^{-1} + B_1 F_1 & -A_{12} \\ (B_2 - K_r B_1)F_2 & (B_2 - K_r B_1)F_1 & \Phi_{22}^{-1} + K_r A_{12} \end{bmatrix} \\ &= \det \begin{bmatrix} \Phi_{22}^{-1} + B_2 F_2 & -A_{21} + B_2 F_1 & -\Phi_{22}^{-1} \\ -A_{12} + B_1 F_2 & \Phi_{11}^{-1} + B_1 F_1 & -A_{12} \\ (B_2 - K_r B_1)F_2 & (B_2 - K_r B_1)F_1 & \Phi_{22}^{-1} + K_r A_{12} \end{bmatrix} \\ &= \det \begin{bmatrix} \Phi_{11}^{-1} + B_1 F_1 & -A_{12} + B_1 F_2 & -A_{12} \\ -A_{21} + B_2 F_1 & \Phi_{22}^{-1} + B_2 F_2 & -\Phi_{22}^{-1} \\ (B_2 - K_r B_1)F_1 & (B_2 - K_r B_1)F_2 & \Phi_{22}^{-1} + K_r A_{12} \end{bmatrix} \\ &= \det \begin{bmatrix} \Phi^{-1} + BF & \Phi_a \\ (B_2 - K_r B_1)F & \Phi_{22}^{-1} + K_r A_{12} \end{bmatrix} \\ &= \det[\Phi_{22}^{-1} + K_r A_{12}] \\ &\quad \cdot \det(\Phi^{-1} + HF - \Phi_a(\Phi_{22}^{-1} + K_r A_{12})^{-1}(B_2 - K_r B_1)F) \\ &= \det[\Phi_{22}^{-1} + K_r A_{12}] \det[\Phi^{-1}] \\ &\quad \cdot \det\{I_n + [\Phi B - \Phi\Phi_a(\Phi_{22}^{-1} + K_r A_{12})^{-1}(B_2 - K_r B_1)]F\} \\ &= \det[\Phi_{22}^{-1} + K_r A_{12}] \det[\Phi^{-1}] \\ &\quad \cdot \det\{I_m + F\Phi B - F\Phi\Phi_a(\Phi_{22}^{-1} + K_r A_{12})^{-1}(B_2 - K_r B_1)\} \quad (\text{G.2}) \\ &= \det[\Phi_{22}^{-1} + K_r A_{12}] \det[\Phi^{-1}] \\ &\quad \cdot \det\{I_m + F\Phi B - F_2(\Phi_{22}^{-1} + K_r A_{12})^{-1}(B_2 - K_r B_1)\} \quad (\text{G.3}) \\ &= \det[\Phi_{22}^{-1} + K_r A_{12}] \det[\Phi^{-1}] \det\{I_m + F\Phi B - M_r(s)\}. \quad (\text{G.4}) \end{aligned}$$

We used (G.1) in order to get (G.3) from (G.2). Noting that

since $K_r \in \mathcal{K}_r$, $M_r(s) = 0$ and hence

$$\begin{aligned} \det [sI_{2n-p} - A_{clw}] &= \det [\Phi_{22}^{-1} + K_r A_{12}] \det [\Phi^{-1}] \det \{I_m + F\Phi B\} \\ &= \det [\Phi_{22}^{-1} + K_r A_{12}] \det [\Phi^{-1}] \det \{I_m + \Phi BF\} \\ &= \det [\Phi_{22}^{-1} + K_r A_{12}] \det [\Phi^{-1} + BF]. \end{aligned}$$

This proves the theorem.

APPENDIX H: PROOF OF THEOREM 10

Since $K_r(\sigma) \in \mathcal{K}_r(\sigma)$, $M_r(s)$ tends to zero point-wise in s as

$\sigma \rightarrow \infty$. Then from (G.4),

$$\begin{aligned} \det [sI_{2n-p} - A_{clw}] &= \det [\Phi_{22}^{-1} + K_r(\sigma)A_{12}] \det [\Phi^{-1}] \det \{I_m + F\Phi B - M_r(s)\} \\ &\rightarrow \det [\Phi_{22}^{-1} + K_r(\sigma)A_{12}] \det [\Phi^{-1}] \\ &\quad \cdot \det \{I_m + F\Phi B\} \text{ as } M_r(s) \rightarrow 0 \\ &= \det [\Phi_{22}^{-1} + K_r(\sigma)A_{12}] \det [\Phi^{-1}] \det \{I_m + \Phi BF\} \\ &= \det [\Phi_{22}^{-1} + K_r(\sigma)A_{12}] \det [\Phi^{-1} + BF]. \end{aligned}$$

This proves the theorem.

Trade-offs in Linear Control System Design*

R. H. MIDDLETON†

Unstable poles, non-minimum phase zeros and time delays in an open loop plant impose fundamental constraints on achievable performance in control loops, and also imply restrictions on the bandwidth of the control loop.

Key Words—Control systems, discrete time systems, non-minimum phase systems, unstable systems, time delays

—For some time now, many practitioners and researchers in the control area have been aware that unstable open loop poles, non-minimum phase zeros, and/or time delays make control systems design difficult. In this paper we examine the nature of these difficulties by discussing the results of Freudenberg and Looze (1987, 1988) and Sung and Hara (1988) on integral constraints on sensitivity functions. One of the key conclusions here is a set of rules of thumb, giving limitations on the closed loop bandwidth which are imposed by unstable open loop poles, non-minimum phase zeros and/or time delays.

1. INTRODUCTION

MUCH of the control literature to date has addressed a question of the form "Given a nominal plant and a set of design criteria, design a controller to best achieve these criteria". Some examples of this are quantitative feedback theory (Horowitz, 1963), where the design criteria takes the form of frequency domain performance specifications in the face of plant uncertainty; LQG control (e.g. Kwakernaak and Sivan, 1972); pole-placement (e.g. Middleton and Goodwin, 1990); and H_∞ optimal control (e.g. Francis, 1987).

In this paper a different question is addressed, namely "Given a nominal plant, what fundamental constraints are implied by unstable open loop poles, non-minimum phase zeros and/or time delays?" The main aim of this paper is to answer this question in terms of constraints on the bandwidth of a control system, imposed by unstable open loop poles etc.

Many of the works on H_∞ optimal control have demonstrated certain constraints imposed by unstable open loop poles and/or non-minimum phase zeros. For example, in many of the early

works on H_∞ optimal control using interpolation theory, it was clear that the interpolation constraints, imposed by unstable plant poles and/or zeros, made controller design more difficult (see for example Zames and Francis, 1983). It has also been shown, (e.g. Francis and Zames, 1984) that non-minimum phase zeros dictate a lower bound on the weighted sensitivity function. This conclusion, combined with the ubiquitous case of having a low pass weighting function, suggests that non-minimum phase zeros place an upper limit on the bandwidth (see for example Zames, 1981, Remark following Theorem 4). O'Young and Francis (1986) have discussed some trade-offs in the case where a description of the plant uncertainty is given.

Other works which discuss constraints imposed by unstable open loop poles etc. include the works of Freudenberg and Looze (1988) and Sung and Hara (1988). In these works, integral constraints on the sensitivity and complementary sensitivity functions are presented. These constraints can be clearly seen to be worsened by the addition of unstable poles, non-minimum phase zeros and/or time delays.

This paper unifies the work (in continuous time), of Freudenberg and Looze (1987, 1988) and the discrete time results of Sung and Hara (1988). In addition, time domain results concerning undershoot and overshoot are discussed here. The main contribution here is to argue for some simple rules of thumb, constraining the bandwidth, based on both frequency domain sensitivity functions and time domain arguments.

This paper is organized as follows. In Section 2 we present several time domain constraints on the response of a stable closed loop, feedback system, imposed by unstable open loop poles and/or non-minimum phase zeros. In Section 3, we discuss the implications of these constraints

*Received 7 June 1989, revised 15 December 1989, revised 12 June 1990, received in final form 29 June 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor H. Kimura under the direction of Editor H. Kwakernaak.

†Department of Electrical Engineering and Computer Science, University of Newcastle, N.S.W., Australia.

with particular reference to rise time, settling time, undershoot and overshoot. Presented in Section 4 are several frequency domain sensitivity function constraints which unify the work of Freudenberg and Looze (1988) and Sung and Hara (1988). In addition, we present a new, closely related result, Lemma 4.3. In Section 5 we discuss the implications of these constraints with particular reference to the closed loop bandwidth. This discussion, together with that in Section 3, leads to several design guidelines which are presented in Section 6. These guidelines are illustrated in Section 7 by discussion of the well known inverted pendulum problem.

In this paper the following unified notation afforded by the use of delta operators (see Middleton and Goodwin, 1990) is used. The delta operator notation was chosen for two reasons: (i) This notation allows a unified presentation where both continuous and discrete results are treated within one framework; and (ii) except in the case of slow sampling, the delta operator has superior numerical properties (see for example Middleton and Goodwin, 1986). The generalized notation is:

(Generalized Derivative)

$$\delta x(t) \triangleq \frac{x(t + \Delta) - x(t)}{\Delta} \rightarrow \frac{dx}{dt} \quad (1.1)$$

Given the above definition (1.1), it can be shown that the state transition matrix for the Generalized Linear time invariant system,

$$\delta x(t) = Ax(t)$$

is the following Generalized Exponential:

$$E(A, t) \triangleq (I + A\Delta)^{t/\Delta} \rightarrow e^{At} \quad (1.2)$$

The inverse operation of the generalized derivative can be shown to be the following generalization of integration:

(Lower Riemann Sum)

$$\int_0^t f(\tau) d\tau \triangleq \Delta \sum_{k=0}^{t/\Delta-1} f(k\Delta) \rightarrow \int_0^t f(\tau) d\tau \quad (1.3)$$

Given the above definitions, the following transform definition will be used:

(Delta Transform)

$$F(\gamma) \triangleq \sum_0^\infty E(\gamma, -t) f(t) dt \rightarrow \text{Laplace Transform} \quad (1.4)$$

Note that the variable γ in the Delta Transform is related to the δ operator in much the same way as s is related to $\frac{d}{dt}$, or z is related

to the shift operator, q . Also, $F(\gamma) = \Delta F_2(z)|_{z=1+\Delta\gamma}$, where $F_2(\cdot)$ is the Z transform of $f_k \triangleq f(k\Delta)$.

Given (1.2), it can be seen that the stability region is the circular region, $|1 + \Delta\gamma| < 1$, which can be rewritten as:

(Stability Region)

$$\text{Re} \{\gamma\} + \frac{\Delta}{2} |\gamma|^2 < 0 \rightarrow \text{Left Half Plane} \quad (1.5)$$

The associated stability boundary is given by:

(Stability Boundary)

$$\gamma = \left(\frac{e^{j\omega\Delta} - 1}{\Delta} \right) \rightarrow \text{Imaginary Axis} \quad (1.6)$$

2 TIME DOMAIN CONSTRAINTS

Consider the simple feedback control system shown in Fig. 1, which by use of the unified notation introduced above, includes both the continuous and discrete cases. (For the remainder of this paper, unless otherwise stated, all results apply to both the discrete time case and the continuous time case.)

In Fig. 1, the symbols have the following meaning:

- $G(\gamma)$ plant transfer function
- $C(\gamma)$ controller transfer function
- $y^*(t)$ desired output, i.e. set point for y
- $e(t)$ error signal
- $u(t)$ system input
- $y(t)$ plant output.

We shall also use $L(\gamma)$ to denote the open loop transfer function

$$L(\gamma) = G(\gamma)C(\gamma) \quad (2.1)$$

(where we assume $L(\gamma)$ is rational, except for a possible multiplicative time delay in continuous time).

We also use $S(\gamma)$, $T(\gamma)$ to denote the sensitivity and complementary sensitivity functions respectively;

$$S(\gamma) \triangleq \frac{1}{1 + L(\gamma)} \quad (2.2)$$

and

$$T(\gamma) \triangleq 1 - S(\gamma) = \frac{L(\gamma)}{1 + L(\gamma)} \quad (2.3)$$

We shall say the closed loop system is "stable" if and only if all of $T(\gamma)$, $S(\gamma)$, $C(\gamma)S(\gamma)$ and

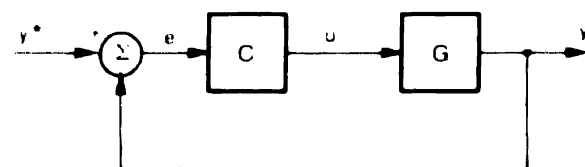


FIG. 1.

$G(\gamma)S(\gamma)$ are analytic for all γ outside the stability region (1.5).

We then have the following results:

Lemma 1. ("Unstable" open loop poles). Let $e(t)$, $y(t)$ denote the responses for $y^*(t)$ a unit step, and suppose there is an open loop pole at $\gamma = p$ (p outside the stability region) i.e. $L(p) = \infty$. Then, for any stable closed loop system:

$$(a) \quad \int_0^\infty E(p, -t)e(t) dt = 0 \quad (2.4)$$

and (b)

$$\int_0^\infty E(p, -t)y(t) dt = \frac{1 + \Delta p}{p} \quad (2.5)$$

Proof. Let $Y^*(\gamma)$ be the transform of $y^*(t)$, (i.e. $Y^*(\gamma) = \frac{(1 + \Delta\gamma)}{\gamma}$). Then

$$E(\gamma) = S(\gamma)Y^*(\gamma) \quad (2.6)$$

Since the closed loop is stable, $\gamma = p$ is in the region of convergence of $E(\gamma)$, and thus:

$$\int_0^\infty E(p, -t)e(t) dt = E(p) = S(p)Y^*(p) = 0 \quad (2.7)$$

Part (b) follows, since $y^*(t) = 1$, and

$$\int_0^\infty E(p, -t) dt = \frac{1 + \Delta p}{p} \quad (2.8)$$

□□□

Lemma 2. ("Non-minimum phase" zeros). Let $e(t)$, $v(t)$ denote the response for $y^*(t)$ a unit step, and suppose there is a zero at $\gamma = z$ (z outside the stability region) i.e. $L(z) = 0$; then for any stable closed loop system:

$$(a) \quad \int_0^\infty E(z, -t)v(t) dt = 0 \quad (2.9)$$

and (b)

$$\int_0^\infty E(z, -t)e(t) dt = \frac{1 + \Delta z}{z} \quad (2.10)$$

Proof. As for Lemma 1, except that in this case, $Y(z) = 0$. □□□

Remark 1. Lemma 2, part (a) applies also in the case where the plant, G , has an unstable zero at $\gamma = z$, and the control signal is any bounded, integrable function. Thus, in particular, this constraint is independent of whether the controller is linear or nonlinear, time invariant or time varying, etc. □□□

3. INTERPRETATION OF TIME DOMAIN CONSTRAINTS

In this section we discuss the implications of the results in Section 2. We first show that real right half plane poles in the open loop imply overshoot in the closed loop step response.

Lemma 3 (Rise time, overshoot and real RHP poles).

- (a) A stable unit feedback system which has a real, right half-plane open loop pole, must have overshoot in its step response.
(b) The amount of overshoot is related to the "rise time" and the location of the right half plane pole, p , as follows. Define the "rise time", t_r , as:

$$t_r = \sup \left\{ T: y(t) \leq \frac{T}{T} \text{ for } t \in [0, T] \right\} \quad (3.1)$$

(and where in discrete time, the sup is over $T = k\Delta$, $k \in \mathbb{Z}^+$) then the overshoot, $y_{\text{ov}} = \sup \{y(t) - 1\}$, satisfies

$$y_{\text{ov}} \approx \frac{1}{pt_r} [(pt_r - 1)E(p, t_r) + 1] \approx \frac{p(t_r + \Delta)}{2} \quad (3.2)$$

Proof. (a) Follows directly from Lemma 1(a) since for $p \in \mathbb{R}^+$, $E(p, t) > 0$.

(b) From the rise time definition, (3.1),

$$e(t) \geq \left(1 - \frac{t}{t_r}\right); \quad t \in [0, t_r] \quad (3.3)$$

Using (3.3) in the integral equality (2.4) we obtain:

$$-\int_0^\infty E(p, -t)e(t) dt \geq \int_0^{t_r} E(p, -t)\left(1 - \frac{t}{t_r}\right) dt \quad (3.4)$$

From (3.4) and the definition of the overshoot, y_{ov} , it follows that

$$y_{\text{ov}} \int_0^\infty E(p, -t) dt \geq \int_0^{t_r} E(p, -t)\left(1 - \frac{t}{t_r}\right) dt \quad (3.5)$$

which leads to (3.2) since

$$\begin{aligned} & \left(\int_0^{t_r} E(p, -t)\left(1 - \frac{t}{t_r}\right) dt \right) / \left(\int_0^\infty E(p, -t) dt \right) \\ &= \frac{1}{pt_r} [(pt_r - 1)E(p, t_r) + 1]. \end{aligned} \quad (3.6)$$

□□□

Thus it is seen that long rise times (i.e. "slow" closed loop response) in a system with real unstable open loop poles implies large overshoot in a unity feedback configuration. This differs from the case of an open loop stable system

where we would normally expect short rise times to give large overshoot. The result here can be thought of as a "demand" for "fast action" imposed by unstable poles in the open loop.

Using the standard rule of thumb (e.g. Middleton and Goodwin, 1990), the rise time, t_r , can be related to the bandwidth, ω_B , (in rad s^{-1}) via:

$$t_r \omega_B \approx 2.3 \quad (3.7)$$

where the bandwidth, ω_B , is defined to be the angular frequency where the magnitude of the complementary sensitivity function is -3 dB , i.e.

$$\left| T \left(\frac{e^{j\omega_B \Delta} - 1}{\Delta} \right) \right| = \frac{1}{\sqrt{2}} \quad (3.8)$$

We assume that a unique frequency ω_B within the range $\left(0, \frac{\pi}{\Delta}\right)$ exists such that (3.8) is satisfied. Note that the rule of thumb, (3.7) is approximately true for a wide range of systems, $T(y)$, which have a "reasonable" step response. Cases where this rule breaks down may include systems with lightly damped complex poles in the closed loop (in which case ω_B may not be well defined).

From (3.7) we see that to be able to keep the overshoot y_{os} below about 20% we require

$$\omega_B \leq \frac{6p}{1 - 2.5p\Delta} \quad (3.9)$$

In the next lemma we show that real, right half plane zeros imply undershoot in the step response, and that to keep the undershoot small, a low bandwidth is required.

Lemma 4. (Settling time, undershoot and real RHP zeros).

- A stable closed loop system which has a real, right half-plane, open loop zero must have undershoot in its step response.
- The amount of undershoot is related to the settling time and the location of the zero, z , as follows. Defining the "settling time" t_s as:

$$t_s = \inf_T \{ T : y(t) \geq 0.9 \text{ for } t \in [T, \infty) \} \quad (3.10)$$

(and where in discrete time, the inf is over $T = k\Delta$, $k \in \mathbb{Z}^+$) then the undershoot, $y_{us} \triangleq \sup_t \{-y(t)\}$ satisfies

$$y_{us} \geq \frac{0.9}{E(z, t_s) - 1} \quad (3.11)$$

Proof (similar to Lemma 3)

$$Y(z) = 0 = \sum_0^\infty E(z, -t)y(t) dt \quad (3.12)$$

With the settling time, t_s , as defined in (3.10) and the undershoot as defined above, we have from (3.12):

$$y_{us} \sum_0^{t_s} E(z, -t) dt \geq \sum_{t_s}^\infty 0.9 E(z, -t) dt \quad (3.13)$$

From (3.13) it is clear that for a given z , the undershoot must become large as t_s becomes small. In particular (3.11) follows from (3.13).

▽▽▽

If we assume, as is usually desirable, that the settling time is approximately equal to the rise time, and using the rule of thumb relating the bandwidth, ω_B , and the rise time, t_r , (3.7), we note that to keep the minimum undershoot given by (3.12) below about 10% we require

$$\omega_B \leq \frac{1}{\Delta} \ln(1 + \Delta z) \approx z \left(1 - \frac{\Delta z}{2}\right) \quad (3.14)$$

Thus we see that unstable zeros place an upper limit on the bandwidth that can be used.

Remark 2. As discussed in Remark 1 the above conclusion is, in fact, independent of the type of controller used; i.e. (3.10) holds for any (Non-Linear, Time Varying) control signal generated in open or closed loop etc. provided only that the input step response is bounded.

▽▽▽

Remark 3. The above results appear to contradict the results of Clark (1962), and Vidyasagar (1986) where system "undershoot" occurs if and only if there are an odd number of real right half plane zeros. This paradox arises due to different definitions of "undershoot". In the above two references, a system is said to undershoot if the initial sign of $y(t)$ is negative.

▽▽▽

The following lemma describes the trade-offs inherent in systems with a real right half plane pole and a time delay.

Lemma 5 (Real RHP pole and time delay). A stable unity feedback system with a time delay, τ , and a real right half plane open loop pole at $y = p$, must have overshoot, y_{os} , which satisfies

$$y_{os} \geq E(p, \tau) - 1 \geq p\tau \quad (3.15)$$

Proof. Follows directly from Lemma 1 on noting that $y(t) = 0$ for $t < \tau$.

▽▽▽

From (3.15) it is clear that to be able to keep the overshoot below 20% we require:

$$p\tau \leq 0.2 \quad (3.16)$$

In discrete time, τ will be at least Δ in practice, and so (3.16) suggests:

$$\Delta \leq \frac{0.2}{p}. \quad (3.17)$$

The following lemma discusses the case where we have real right half plane poles and zeros in the open loop transfer function.

Lemma 6 (Real RHP poles and zeros). Suppose a stable, unity feedback system has a real right half plane open loop pole at $\gamma = p$, and a real half plane zero at $\gamma = z \neq p$; then (a) if $p \leq z$, the overshoot, satisfies:

$$y_{os} \geq \frac{1}{\frac{z(1+\Delta p)}{p(1+\Delta z)} - 1}. \quad (3.18)$$

(b) If $p > z$, the undershoot satisfies

$$y_{us} \geq \frac{1}{\frac{p(1+\Delta z)}{z(1+\Delta p)} - 1}. \quad (3.19)$$

Proof. (a) From Lemma 1 and Lemma 2 we have

$$\int_0^{\infty} (E(p, -t) - E(z, -t))y(t) dt = \frac{1+\Delta p}{p}. \quad (3.20)$$

The result follows since

$$E(p, -t) \geq E(z, -t) \quad \text{for } p > z.$$

(b) Follows similarly to (a). $\square \square \square$

We note from part (a) that for $0 < p < z$, an overshoot less than 20% can only be achieved if:

$$\frac{z}{1+\Delta z} \geq 6 \frac{p}{1+\Delta p}. \quad (3.21)$$

(Note that for $p\Delta > 0.2$, there is no $z > 0$ such that (3.21) is satisfied.)

Conversely, for $0 < z < p$ an undershoot of less than 10% can only be achieved if

$$\frac{p}{1+\Delta p} \geq 11 \frac{z}{1+\Delta z}. \quad (3.22)$$

(Note that for $\Delta z > 0.1$, there is no $p > 0$ such that (3.22) is satisfied.) We would argue, from Lemma 3, Lemma 4 and (3.22) that in fact, it is very difficult to achieve satisfactory performance if $z < p$. To see this, we will assume Δ is small, in which case (3.22) becomes:

$$p \geq 11z. \quad (3.23)$$

From (3.2), to be able to keep the overshoot

below 20%, we require

$$t_r \leq \frac{2 \times 0.2}{p} \leq \frac{0.4}{11z}. \quad (3.24)$$

From (3.11), to be able to keep the undershoot below 10% we require:

$$zt_r \geq 2.3. \quad (3.25)$$

From (3.25) and (3.24) we see that for $0 < z < p$, less than 20% overshoot, and less than 10% undershoot, we must have

$$t_s \geq 63t_r. \quad (3.26)$$

Thus the system must rise quickly but settle slowly. Although this wide discrepancy in rise time and settling time is possible, we suggest that it is almost always highly undesirable, and thus we suggest that the case $0 < z < p$ is very difficult to control.

Similar arguments do *not* lead to the same conclusion in the case where $0 < p < z$. In particular, provided (3.21) is satisfied, it appears that control with reasonable overshoot, undershoot, rise and settling times is possible.

4. FREQUENCY DOMAIN SENSITIVITY FUNCTION CONSTRAINTS

It can be seen above that there are fundamental limits on the bandwidth imposed by the locations of the poles and zeros of the system. In Section 3 the bandwidth limitations were motivated by time domain arguments. To gain further insight into the constraints, we next view them from a frequency domain perspective. To do this the sensitivity functions, $S(\gamma)$ and $T(\gamma)$, defined in (2.2), (2.3) are considered.

Having S small reduces the closed loop sensitivity to disturbances and plant variations whilst having T small reduces the closed loop sensitivity to measurement errors. However, both sensitivity functions cannot be simultaneously small since $S + T = 1$.

It will be seen below that there are integral constraints on these sensitivity functions. The constraints presented here are unifications of the results in Freudenberg and Looze (1988) and Sung and Hara (1988) which are based on the work of Bode (1945).

Lemma 7 (Sensitivity function and unstable poles). Provided the open loop transfer function has relative degree ≥ 1 (discrete time or continuous time with a pure time delay) or > 1 (Continuous Time), and the closed loop system is stable, then the sensitivity function, $S(\gamma)$, satisfies the following integral constraint:

$$\int_0^{\pi/\Delta} \log \left| S \left(\frac{e^{j\omega\Delta} - 1}{\Delta} \right) \right| d\omega = \pi \sum_{i=1}^N p_i' \quad (4.1)$$

where $S\left(\frac{e^{j\omega\Delta}-1}{\Delta}\right)$ is the magnitude of the frequency response of the sensitivity function at ω rad sec⁻¹, U is the set of integers, i , such that p_i (the i th open loop pole) is unstable, and where p_i^* is the continuous counterpart of the i th unstable pole, p_i , i.e.

$$p_i^* \triangleq \frac{1}{\Delta} \log(1 + \Delta p_i). \tag{4.2}$$

Proof. For the continuous time case, see Freudenberg and Looze (1987). The discrete time case follows directly from Sung and Hara (1988), Theorem 2; by using the change of variable $\omega\Delta = \phi$, and by translating from the z plane in Sung and Hara (1988) to the γ plane where $\gamma = \frac{z-1}{\Delta}$. ▽▽▽

Lemma 8 (Sensitivity function, unstable poles and zeros). Suppose a stable closed loop system has open loop poles, p_i ($i \in U$) which are outside the stability region, and zeros, z_i ($j \in U'$) which are outside the stability region, then for all $j \in U'$:

$$\pi \log |B_p^{-1}(z_i)| = \int_{-\pi/\Delta}^{\pi/\Delta} \frac{X(z_j)}{|e^{j\omega\Delta}-1|} \log \left| S\left(\frac{e^{j\omega\Delta}-1}{\Delta}\right) \right| d\omega \tag{4.3}$$

where

$$X(\gamma) \triangleq \operatorname{Re}\{\gamma\} + \frac{\Delta}{\gamma} |\gamma|^2 \tag{4.4}$$

and $B_p(\gamma)$ is the discrete time delta operator Blaschke product of the poles:

$$B_p(\gamma) \triangleq \prod_{i \in U} \left| \frac{1}{1 + \Delta p_i^*} \left(\frac{p_i - \gamma}{\bar{p}_i - \gamma} \right) \right| \tag{4.5}$$

where

$$\bar{p}_i \triangleq -p_i^*/(1 + \Delta p_i^*). \tag{4.6}$$

Proof. See Freudenberg and Looze (1987) for the continuous case, and Sung and Hara (1988) for the discrete time case, where $\omega\Delta = \phi$ and $\gamma = \frac{z-1}{\Delta}$. ▽▽▽

Lemma 9 (Unstable zeros, time delays and the complementary sensitivity function). Provided the open loop transfer function $L(\gamma)$ has at least one integrator and provided the closed loop is stable, then the complementary sensitivity function, T , satisfies the following integral

constraint

$$\int_0^\infty \frac{1}{\omega} \operatorname{sinc}\left(\frac{\omega\Delta}{2}\right) \log \left| T\left(\frac{e^{j\omega\Delta}-1}{\Delta}\right) \right| d\omega \\ H_0^{-1} + \pi \sum_{i \in U'} \left(\frac{1}{z_i} + \frac{\Delta}{2} \right) + \frac{\pi}{2} \tau \tag{4.7}$$

where : U' is the set of zeros i such that the i th zero, z_i , is outside the stability region; τ is the system time delay (defined to be $r\Delta$ in discrete time, where r is the relative degree); $\operatorname{sinc}(x)$ is $\sin(x)/x$; and H_0 is the system velocity constant,

$$H_0 \triangleq \lim_{\gamma \rightarrow 0} (\gamma L(\gamma)) \neq 0. \tag{4.8}$$

Proof. This is a dual result of Lemma 7 which appears to have been unnoticed previously. See Appendix A for a proof. ▽▽▽

Lemma 10 (Complementary sensitivity function). For any stable closed loop system, we have for all $i \in U$

$$\pi |p_i| \tau \log |B_z^{-1}(p_i)| = \int_{-\pi/\Delta}^{\pi/\Delta} \frac{X(p_i)}{|e^{j\omega\Delta}-1|} \times \log \left| T\left(\frac{e^{j\omega\Delta}-1}{\Delta}\right) \right| d\omega \tag{4.9}$$

where τ is as defined in Lemma 9 and $B_z(\gamma)$ is the discrete time delta operator Blaschke product of the zeros

$$B_z(\gamma) \triangleq \prod_{i \in U'} \left| \frac{1}{1 + \Delta z_i^*} \left(\frac{z_i - \gamma}{\bar{z}_i - \gamma} \right) \right|$$

Proof. Dual of Lemma 8, or, see Freudenberg and Looze (1988) for the continuous case, or Sung and Hara (1988) for the discrete case. ▽▽▽

In the following section we discuss some of the implications of the above lemmata.

5 INTERPRETATION OF FREQUENCY DOMAIN CONSTRAINTS

In the previous section we gave four results giving fundamental constraints on the sensitivity functions. In each of the four cases, the amplitude of the frequency response satisfies an integral equality. The discussion in this section will centre on lower bounds on peaks in the sensitivity function. To aid our discussion we make the following definition. For any stable $G(\gamma)$ we define:

$$\|G(\cdot)\|_\infty = \sup_{\omega \in \mathbb{R}} \left| G\left(\frac{e^{j\omega\Delta}-1}{\Delta}\right) \right|. \tag{5.1}$$

Lemma 11 (Unstable pole(s) and Time Delays). Consider any stable closed loop system where the open loop system has an unstable pole at $\gamma = p_i$ and a time delay of τ , then

$$\|T(\cdot)\|_\infty \geq |e^{p_i \tau}|. \quad (5.2)$$

Proof. Follows directly from Lemma 10 since:

$$\int_{-\pi/\Delta}^{\pi/\Delta} \frac{X(p_i)}{\left| \frac{e^{j\omega\Delta} - 1}{\Delta} - p_i \right|^2} d\omega = \pi. \quad (5.3)$$

From Lemma 11, we note that to be able to achieve a peak in $T(\gamma)$ of less than about 1.2, we require

$$\forall i, \tau \operatorname{Re}\{p_i\} \leq 0.2. \quad (5.4)$$

Note that this agrees with the conclusion from time domain arguments in (3.15).

Lemma 12 (Unstable pole and unstable zero). Consider any stable closed loop system where the open loop plant has unstable poles and non-minimum phase zeros; then:

$$\|T(\cdot)\|_\infty \geq |B_p^{-1}(p_i)| \quad (5.5)$$

$$\|S(\cdot)\|_\infty \geq |B_p^{-1}(z_i)| \quad (5.6)$$

Proof. Follows directly from Lemmata 8 and 10 using (5.3). See also Freudenberg and Looze (1988) and Sung and Hara (1988). \square

If we take, for example, the case where we have a single real pole, p , and a single real zero, z , Lemma 12 gives:

$$\|T(\cdot)\|_\infty \geq \left| \frac{p + z + \Delta pz}{p - z} \right| \quad (5.7)$$

and

$$\|S(\cdot)\|_\infty \geq \left| \frac{p + z + \Delta pz}{p - z} \right| \quad (5.8)$$

(See also Freudenberg and Looze, 1987.)

Thus, to be able to keep the peaks in T and S less than about 1.5, we require either:

$$z \leq \frac{p}{5 + 2\Delta p} \leq 0.2p \quad (5.9)$$

or

$$z \geq \frac{p}{0.2 - 0.4\Delta p} \geq 5p. \quad (5.10)$$

(See also Lemma 6 for similar time domain constraints.)

We next consider the integral constraint (4.3). In doing this, we first give approximate bounds on the sensitivity function S . From experience, the sensitivity function has the general form

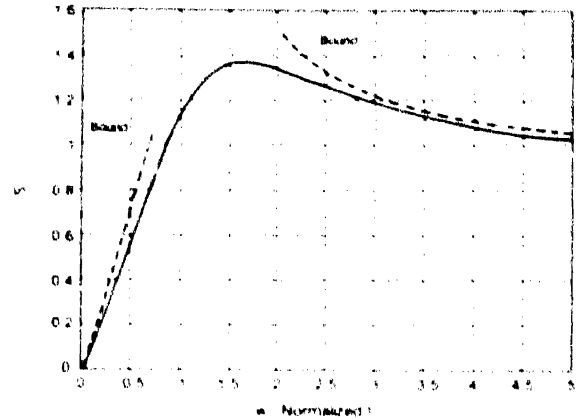


Fig. 2

shown in Fig. 2 where

$$\left| S\left(\frac{e^{j\omega\Delta} - 1}{\Delta}\right) \right| \triangleq \hat{S}(\omega),$$

and the normalized angular frequency, ω is $\frac{\omega}{\omega_n}$.

Upper bounds on $\hat{S}(\omega)$ are typically

$$\hat{S}(\omega) \approx \frac{1.5\omega}{\omega_n} \quad \text{for} \quad \frac{\omega}{\omega_n} < 0.75 \quad (5.11)$$

$$\hat{S}(\omega) \approx 1 + \frac{2\omega_n^2}{\omega^2} \quad \text{for} \quad \frac{\omega}{\omega_n} > 2. \quad (5.12)$$

(Note that if $L(\gamma)$ is strictly proper, and includes at least one integrator, then $S(0) = 0$ and $S(\infty) = 1$.)

We then have the following result.

Lemma 13 (Unstable poles, bandwidth and sensitivity). For any system which is closed loop stable, has relative degree greater than 1, and, (5.11) and (5.12) are satisfied, then

$$\|S(\cdot)\|_\infty \geq \exp\left(\frac{\pi}{2\omega_n \Delta} \sum p_i' + 0.32\omega_n \Delta\right) \quad (5.13)$$

Proof. Using (4.3), we have

$$\begin{aligned} \pi \sum_{i=1}^N p_i' &= \int_0^{\pi/\Delta} \log \hat{S}(\omega) d\omega \\ &= \int_0^{0.75\omega_n} \log \hat{S}(\omega) d\omega \\ &\quad + \int_{0.75\omega_n}^{2.75\omega_n} \log \hat{S}(\omega) d\omega \\ &\quad + \int_{2.75\omega_n}^{\pi/\Delta} \log \hat{S}(\omega) d\omega. \end{aligned} \quad (5.14)$$

Using the bounds given in (5.11), (5.12) we have

$$\int_{0.75\omega_B}^{2.75\omega_B} \log \tilde{S}(\omega) d\omega \geq \pi \sum_{i=1}^N p'_i + 0.64\omega_B^2\Delta. \tag{5.15}$$

▽▽▽

Note that this type of argument, namely, “forcing” \tilde{S} to be “small” in some range of frequencies necessarily implies that it is large elsewhere has been used by several other authors. For example, Francis and Zames (1984, Section V) have shown that \tilde{S} can be made arbitrarily small over any interval, but only at the “cost” (when there are non-minimum phase zeros present) of arbitrarily large sensitivity outside the interval. Freudenberg and Looze (1988) consider several cases such as (i) keeping the sensitivity small over an interval, and (ii) “Relative degree” type assumption on $L(\tilde{S})$ (e.g. $|L(j\omega)| \leq \frac{k}{\omega^r}$; $\omega > \omega_1$).

From Lemma 13, if the lower limit on the peak in $\tilde{S}(\omega)$ is to be not more than 1.4, then it is approximately required that

$$\omega_B(1 + \omega_B\Delta) \leq 5 \sum_{i=1}^N p'_i. \tag{5.16}$$

Thus [as in (3.8)] unstable poles place a lower limit on the bandwidth that should be used.

Also note from (5.16) that good sensitivity requires $\omega_B\Delta$ to be small compared with 1, i.e. fast sampling is generally desirable.

Next, to investigate the corresponding constraint, Lemma 9, for zeros, complementary sensitivity function $T(\gamma)$ will be used. To do this, it is first noted that the general shape of the function $\left|T\left(\frac{e^{j\omega\Delta}-1}{\Delta}\right)\right|$ is as shown in Fig. 3,

where $v = \frac{\Delta}{2} \cot\left(\frac{\omega\Delta}{2}\right) \approx \omega^{-1}$, and v (normalized) is $v\omega_B$.

Approximate bounds for $\tilde{T}(v)$ are (from

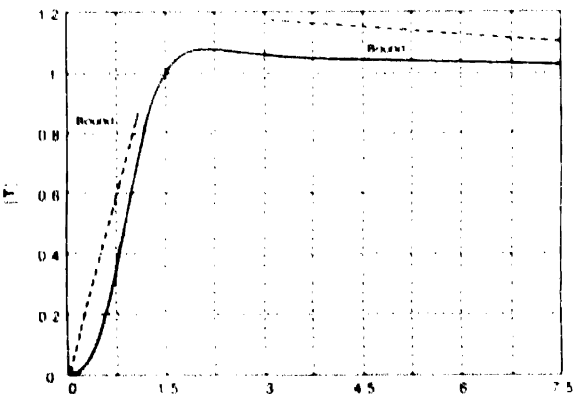


FIG. 3

experience) as follows:

$$\tilde{T}(v) \leq 1.2v\omega_B \quad \text{for } v\omega_B < 0.75 \tag{5.17}$$

$$\tilde{T}(v) \leq \frac{v^2\omega_B^2 + 36}{v^2\omega_B^2 + 30} \quad \text{for } v\omega_B > 2. \tag{5.18}$$

It is also assumed that at low frequencies L falls by at least 20 dB per decade. Hence

$$|H_0| > \omega_B. \tag{5.19}$$

The following result is then obtained.

Lemma 14 (Unstable zeros, delays, bandwidth and sensitivity). For any system that is closed loop stable, and which satisfies the bounds in (5.17) to (5.19), then

$$\|T(\cdot)\|_\infty \geq 0.25 \exp\left(2\omega_B\left[\frac{\tau}{2} + \sum_{i=1}^N \left(\frac{1}{z_i} + \frac{\Delta}{2}\right)\right]\right). \tag{5.20}$$

Proof. Let

$$\Omega \triangleq \frac{\tau}{2} + \sum_{i=1}^N \left(\frac{1}{z_i} + \frac{\Delta}{2}\right). \tag{5.21}$$

Using the bounds (5.19) and (5.18) on the integral (4.7) and splitting the interval of integration gives

$$\begin{aligned} \int_{1/2\omega_B}^{2\omega_B} \log \tilde{T}(v) dv &\geq -\frac{\pi}{2\omega_B} + \pi\Omega \\ &\quad - \int_0^{1/2\omega_B} \log(1.2\omega_B v) dv - \int_{2\omega_B}^\infty \\ &\quad \times \log\left(\frac{v^2\omega_B^2 + 36}{v^2\omega_B^2 + 30}\right) dv. \end{aligned} \tag{5.22}$$

Evaluating the integrals on the right hand side of (5.22) gives

$$\int_{1/2\omega_B}^{2\omega_B} \log \tilde{T}(v) dv \geq -\frac{2.11}{\omega_B} + \pi\Omega \tag{5.23}$$

and the result follows. ▽▽▽

Thus to keep the peak in $\tilde{T}(v)$ below about $\sqrt{2}$, we require

$$\left(\frac{\tau}{2} + \sum_{i=1}^N \left(\frac{1}{z_i} + \frac{\Delta}{2}\right)\right)\omega_B \leq 0.9. \tag{5.24}$$

Thus we note that unstable zeros and/or time delays give an upper bound on the bandwidth which should be used. In the next section, we bring together the results of the previous sections, particularly Sections 3 and 5 in the form of several design guidelines.

6. DESIGN GUIDELINES

6.1. “Avoid using Controllers which have unstable poles or zeros, or which have a time delay”.

We note from all the constraints presented so far that adding a time delay, unstable poles

and/or unstable zeros necessarily makes the sensitivity/undershoot/overshoot constraints worse. If at all possible this situation should be avoided. In cases where a plant can only be stabilized by a controller which is unstable and/or has unstable zeros, alternative solutions should be sought. Possible solutions are the use of additional control actuators, the use of additional state variable measurements and/or plant redesign.

This guideline has been discussed by other authors such as Freudenberg and Looze (1988) and Zames and Francis (1983).

6.2. "When implementing Digital Control laws, do not use slow sampling". In particular we suggest:

$$\omega_s \triangleq \frac{2\pi}{\Delta} \geq 10\omega_B \quad (6.1)$$

Again note that the design tradeoffs become more difficult as Δ increases. Consider for example (5.16). A necessary condition for (5.16) to be satisfied is $\omega_B \Delta < 1$, i.e. $\omega_s > 2\pi\omega_B$. Note also that for a proper, discrete time system, the minimum time delay is Δ , and hence guideline 6.1 suggests Δ be small.

Note that this "rule of thumb" has been suggested in many sampled data control texts; however, few have considered the sensitivity implications of violating this suggestion.

6.3. "If the open loop plant plus controller has unstable poles, the bandwidth of the feedback loop should satisfy:

$$\omega_B(1 - \omega_B \Delta) \geq \frac{5}{\Delta} \sum_{i=1}^n \log(1 + \Delta p_i) \quad (6.2)$$

This guideline is motivated directly from (5.16) and (3.8). (Note that (3.8) leads directly to

$$\begin{aligned} \frac{1}{\Delta} \log(1 + \Delta p) &\leq \frac{1}{\Delta} \log\left(\frac{6 + 3.5\omega_B \Delta}{6 + 2.5\omega_B \Delta}\right) \\ &\approx \frac{1}{6} \omega_B (1 - \omega_B \Delta) \end{aligned}$$

This guideline seems to be less well known than the other guidelines in this section, though Horowitz (1963) seemed to be aware of this constraint.

6.4. "If the open loop plant plus controller has unstable zeros and/or time delays, the bandwidth should satisfy:

$$\omega_B \leq \left(\frac{1}{2} + \sum_{i=1}^n \left(z_i^{-1} + \frac{\Delta}{2} \right) \right)^{-1} \quad (6.3)$$

This guideline comes directly from (5.24) and (3.13). (Note that for a single real RHP zero, and no time delay, (6.3) and (3.13) agree in terms up to

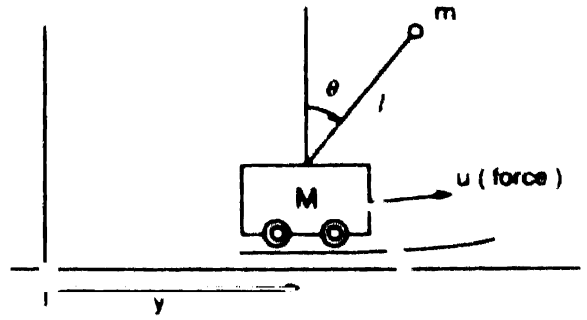


FIG. 4

order Δz^2 .) This guideline is in agreement with the Remark in Zames (1981) following Theorem 4.

Also note that guidelines 6.3 and 6.4 may be mutually contradictory. In this case, alternative approaches to the design problem (as discussed after guideline 6.1, and in the following example) are suggested.

7. INVERTED PENDULUM EXAMPLE

Consider an inverted pendulum system of the type illustrated in Fig. 4.

Using Lagrangian mechanics (see for example Kibble, 1973), a non-linear state space model for this system can be shown to be:

$$\begin{aligned} y &= \frac{1}{\left(\frac{M}{m} + \sin^2 \theta\right)} \\ &\times \left[\frac{u}{m} + \dot{\theta}^2 \ell \sin \theta - g \sin \theta \cos \theta \right] \quad (7.1) \end{aligned}$$

$$\begin{aligned} \dot{\theta} &= \frac{1}{\ell \left(\frac{M}{m} + \sin^2 \theta\right)} \\ &\times \left[-\frac{u}{m} \cos \theta - \dot{\theta}^2 \ell \cos \theta \sin \theta \right. \\ &\quad \left. + \left(1 + \frac{M}{m}\right) g \sin \theta \right] \quad (7.2) \end{aligned}$$

where g denotes the acceleration due to gravity.

A linearized model for this system about the origin is readily seen to be:

$$\begin{aligned} \frac{d}{dt} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{-mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{(M+m)g}{M\ell} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \\ &+ \begin{pmatrix} 0 \\ \frac{1}{M} \\ 0 \\ -\frac{1}{M\ell} \end{pmatrix} u \quad (7.3) \end{aligned}$$

$$y = (1 \quad 0 \quad 0 \quad 0) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \tag{7.4}$$

where $(x_1 \ x_2 \ x_3 \ x_4) = (y, \dot{y}, \theta, \dot{\theta})$.

It is relatively easy to show that the above state space model is both controllable and observable (provided $M, m > 0$) and so a wide range of controller designs are possible. The transfer function from u to y is

$$\frac{Y(s)}{U(s)} = \frac{K(s-b)(s+a)}{s^2(s-a)(s+a)} \tag{7.5}$$

where

$$K = \frac{1}{M} \tag{7.6}$$

$$b = +\sqrt{\frac{g}{\ell}} \tag{7.7}$$

and

$$a = \sqrt{\frac{(M+m)g}{M\ell}} \tag{7.8}$$

Note that, in view of guidelines (6.3) and (6.4) this transfer function is impractical to work with. Firstly it is both unstable and non-minimum phase. Secondly the unstable zero occurs at a lower frequency than the unstable pole. Hence, guidelines (6.3) and (6.4) cannot be both satisfied.

One can still design a controller for the above case, however, the above considerations indicate that the controller will have very poor performance. Consider the situation where $\ell = 1\text{m}$, $g = 10\text{ms}^{-2}$, $M = m = 0.5\text{kg}$. In this case $K = 2$, $b = \sqrt{10}$ and $a = \sqrt{20}$.

Thus, from (5.7) and (5.8) it is clear that any controller design which stabilizes this system will have sensitivities which satisfy

$$\|T(\cdot)\|_1 \geq 5.828$$

$$\|S(\cdot)\|_1 \geq 5.828.$$

Also, from Lemma 6 it can be shown that the undershoot must exceed 241%.

This is clearly unacceptable from a practical point of view. The reader may gain some appreciation of the difficulty of the control problem when only position is measured, by trying to balance a broom with both eyes shut!

In view of the high sensitivity described above, a better option is to consider this system as a single input two output system, i.e. to seek additional state measurements. Suppose, for example, we retain $y_1 = y$ but introduce an additional measurement y_2 , which is the angle of the pendulum, θ .

Let $G_1 = \frac{N_1}{d}$, $G_2 = \frac{N_2}{d}$ denote the transfer

functions from u to y_1 and y_2 respectively:

$$\frac{N_1(s)}{d(s)} = \frac{2(s^2 - 10)}{s^2(s^2 - 20)} \tag{7.9}$$

and

$$\frac{N_2(s)}{d(s)} = \frac{-2s^2}{s^2(s^2 - 20)}. \tag{7.10}$$

Note that the double pole at the origin arising from the acceleration of the cart is unobservable in y_2 . However, we can easily design feedback from y_2 to u so as to shift the other two poles well into the left half plane. Guideline (6.3) suggests $\omega_B = 5\sqrt{20} \doteq 15$.

We thus suggest the following control law:

$$u = v + \frac{170(s + \sqrt{20})}{s + 26} y_2 \\ \triangleq v + K_2(s)y_2. \tag{7.11}$$

The transfer function from v to y_1 then becomes:

$$\frac{y_1(s)}{v(s)} = \frac{2(s + 26)(s^2 - 10)}{s^2(s + \sqrt{20})(s^2 + 21.53s + 223.72)} \tag{7.12}$$

Clearly this gives an easier design problem. Note that the double integrator (which is unobservable from θ) could not be shifted; however, the unstable pole has been shifted by the inner feedback loop. Note also that the unstable zero is still present. In this case design guideline (6.4) suggests that the bandwidth of the outer feedback loop should be $\approx 3\text{rad s}^{-1}$.

The following control law for the position loop is suggested:

$$v = -19.9(y + \sqrt{20})(s^2 + 21.53s + 223.72)(s + 0.16) \\ (s + 26)(s + \sqrt{10})(s^2 + 15.5s + 106.30) \\ \times (y^* - y_1) \triangleq K_1(s)(y^* - y_1). \tag{7.13}$$

To test the sensitivity of the controller given by (7.13) and (7.11), consider the sensitivity functions defined as follows:

$$S(s) = (I + G(s)K(s))^{-1} \tag{7.14}$$

and

$$T(s) = (I + G(s)K(s))^{-1}G(s)K(s) = I - S(s) \tag{7.15}$$

where

$$G(s) \triangleq [G_1(s)G_2(s)]^T$$

and

$$K(s) \triangleq [K_1(s)K_2(s)].$$

Since the main interest is in the behaviour of y , the 1, 1 components of S and T shall be considered, which are:

$$S_{11}(s) = \frac{s^4 + 15.5s^3 + 106.3s^2}{s^4 + 15.5s^3 + 86.4s^2 + 59.7s + 10.1} \tag{7.16}$$

$$T_{11}(s) = \frac{-19.9s^2 + 59.7s + 10.1}{s^4 + 15.5s^3 + 86.4s^2 + 59.7s + 10.1} \quad (7.17)$$

We then find that $\|S_{11}(\cdot)\|_1 = 1.3$, $\|T_{11}(\cdot)\|_1 = 1.2$, the undershoot is 11% and the overshoot is 18%.

Thus, in this example, the design guidelines given suggest that the original problem (i.e. without measurement of θ) is very difficult, while the problem where measurement of the pendulum's angle is available is well known to be easy.

8 CONCLUSIONS

In this paper we have considered both time domain and frequency domain equality constraints which arise in the feedback control of linear systems. Based on these constraints, we have suggested the following guidelines for the selection of the feedback loop bandwidth, ω_R :

$$\omega_R(1 - \omega_R\Delta) \leq 5 \sum_{i \in U} \frac{1}{\Delta} \ln(1 + \Delta p_i) \quad (6.2)$$

(where U is the set of i such that p_i is unstable) and

$$\omega_R \left[\frac{r}{2} + \sum_{i \in U'} \left(\frac{1}{z_i} + \frac{\Delta}{2} \right) \right] \leq 1 \quad (6.3)$$

(where U' is the set of i such that z_i is "unstable", and r is the time delay)

One consequence of the above design rules is that there exist systems which although "controllable" and "observable" (in the usual linear system theoretic sense), cannot be sensibly controlled. In these cases it is suggested that additional measurement transducers/control actuators may be crucial to a successful controller implementation.

REFERENCES

- Bode, H. W. (1945) *Network Analysis and Feedback Amplifier Design*. Van Nostrand, New York.
- Clark, R. N. (1962) *Introduction To Automatic Control Systems*. Wiley, New York.
- Francis, B. A. and G. Zames (1984) On H^∞ optimal sensitivity theory for SISO feedback systems. *IEEE Trans. Aut. Control*, **AC-29**, 9-16.
- Francis, B. A. (1987) *A Course in H^∞ Control Theory*. Springer, Berlin.
- Freudenberg, J. S. and D. P. Looze (1987) A sensitivity trade-off for plants with time delay. *IEEE Trans. Aut. Control*, **AC-32**, 99-104.
- Freudenberg, J. S. and D. P. Looze (1988) *Frequency Domain Properties of Scalar and Multivariable Feedback Systems*. Springer, Berlin.
- Horowitz, I. M. (1963) *Synthesis of Feedback Systems*. Academic Press, New York.
- Kibble, T. W. B. (1966) *Classical Mechanics*. McGraw-Hill, New York.
- Kwakernaak, H. and R. Sivan (1972) *Linear Optimal Control Systems*. Wiley, New York.

- Middleton, R. H. and G. C. Goodwin (1986) Improved finite word length properties in Digital control using delta operators. *IEEE Trans. Aut. Control*, **AC-31**, 1015-1021.
- Middleton, R. H. and G. C. Goodwin (1990) *Digital Control and Estimation: A Unified Approach*. Prentice Hall, Englewood Cliffs, New Jersey.
- O'Young, S. O. and B. A. Francis (1986) Optimal performance and robust stabilization. *Automatica*, **22**, 171-183.
- Sung, H. K. and S. Hara (1988) Properties of sensitivity and complementary sensitivity functions single input-output digital control systems. *Int. J. Control*, **48**, 2429-2439.
- Vidvasagar, M. (1986) On undershoot and nonminimum phase zeros. *IEEE Trans. Aut. Control*, **AC-31**, 440.
- Zames, G. and B. A. Francis (1983) Feedback, minimax sensitivity and optimal robustness. *IEEE Trans. Aut. Control*, **AC-28**, 585-601.
- Zames, G. (1981) Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Trans. Aut. Control*, **AC-26**, 301-320.

APPENDIX A

Proof of Lemma 9

We define

$$r \triangleq \gamma^{-1} \quad (A.1)$$

and then write $T(\gamma)$ as

$$T(r^{-1}) = \frac{1}{1 + H_0(r^{-1})} \quad (A.2)$$

We observe that zeros at γ in the γ plane map to zeros at the origin in the r plane.

We then follow the proof of Lemma 7 in the continuous time case (see Freudenberg and Looze, 1987) using the contour shown in Fig. 5 for the complex r plane. Note that cut-sets are also required for zeros at the origin in the r plane (i.e. zeros at γ in the γ plane), in the discrete time case.

The integral around the total contour $\mathcal{C} = \mathcal{C}_0 + \mathcal{C}_1 + \dots + \mathcal{C}_{N+1}$ is zero. The integral along \mathcal{C}_0 satisfies

$$\lim_{r \rightarrow \infty} \int_{\mathcal{C}_0} \log T(r^{-1}) dr = 2j \int_0^{\pi} \log \left| T \left(\frac{1}{\frac{\Delta}{2} e^{j\theta}} \right) \right| d\theta \quad (A.3)$$

The contribution from each of the contours \mathcal{C}_i , $i = 1, \dots, N$ can be evaluated as

$$\lim_{r \rightarrow \infty} \int_{\mathcal{C}_i} \log T(r^{-1}) dr = 2j\pi \operatorname{Re} \left\{ \frac{1}{z_i} + \frac{\Delta}{2} \right\} \quad (A.4)$$

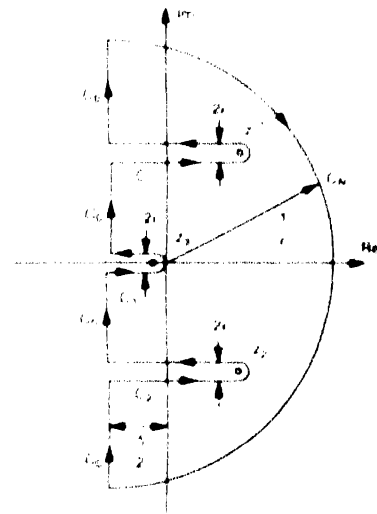


FIG. 5. Contour for complex r -plane used to prove Lemma 7

The contribution from \mathcal{C}_{N+1} can be evaluated as follows.

$$\begin{aligned} \lim_{r \rightarrow 0} \int_{\mathcal{C}_{N+1}} \log T(r^{-1}) \, dr &\approx \lim_{r \rightarrow 0} \int_{\mathcal{C}_{N+1}} \log \left(\frac{1}{1 + \frac{1}{rH_0}} \right) \, dr \\ &= - \lim_{r \rightarrow 0} \int_{\mathcal{C}_{N+1}} \left(\frac{1}{rH_0} \right) \, dr \\ &= - \frac{j\pi}{H_0}. \end{aligned} \tag{A 5}$$

The final term needed for the continuous time case is a cut set to allow for the singularity of $\log T$ caused by the time delay, $e^{-\gamma}$. This time delay causes a singularity at $r = 0$, i.e. $\gamma = \infty$. Thus we take a semi-circle, \mathcal{C}_r , of radius r about the

origin in the r -plane. On this semi-circle we have

$$\begin{aligned} \int_{\mathcal{C}_r} \log T(r^{-1}) \, dr &= \int_{\mathcal{C}_r} \log (e^{-\gamma'}) \, dr \\ &= \int_{\mathcal{C}_r} -\frac{\gamma}{r} \, dr \\ &= -j\pi\gamma. \end{aligned} \tag{A 6}$$

The result then follows from (A.3) to (A.6) on noting that the zeros, z_i , in (A.4) occur in conjugate pairs, and by changing from the variable ν in (A.3) to ω where $\nu = \frac{\Delta}{2} \cot \left(\frac{\omega \Delta}{2} \right)$. □□□

On the Structure of Suboptimal H^∞ Controllers in the Sensitivity Minimization Problem for Distributed Stable Plants*

HITAY ÖZBAY† and ALLEN TANNENBAUM‡

The structure of all the suboptimal H^∞ controllers in the sensitivity minimization are described for a class of distributed plants. For stable plants with continuous transfer functions finite dimensional suboptimal controllers can be obtained by approximating the infinite dimensional part of the optimal controller.

Key Words—Distributed parameter systems, H^∞ -optimal control, sensitivity minimization, suboptimal control, finite dimensional controllers, time delays

Abstract—In this paper we consider the H^∞ sensitivity minimization problem for SISO distributed stable systems. We elucidate the structure of all the suboptimal H^∞ controllers for stable distributed plants with invertible outer parts and rational weights. We identify the finite and infinite dimensional parts of the controller. The conditions under which one can obtain a finite dimensional suboptimal controller by approximating the infinite dimensional parts of the optimal controller are discussed. The case where plant transfer function is continuous on the boundary is also discussed.

1. INTRODUCTION

IN THIS paper we consider the one block H^∞ sensitivity minimization problem for SISO infinite dimensional systems. Our main purpose is to develop a method for obtaining finite dimensional suboptimal H^∞ controllers. We use the results of Foias and Tannenbaum (1989) to describe the structure of all the suboptimal H^∞ controllers. In the case of rational weights and distributed stable plants with invertible outer parts we will be able to identify the finite and infinite dimensional parts of the controller. A natural way of obtaining a finite dimensional suboptimal controller is to approximate the infinite dimensional part of the optimal controller.

It is known that H^∞ -optimal controllers for finite dimensional plants with rational weights are finite dimensional. Therefore, another way to obtain a finite dimensional suboptimal controller is to approximate the original plant with a finite dimensional system, compute the optimal controller for this approximate system, and then check whether this controller is suboptimal for the original plant. However, it is obvious that there is no guarantee that the optimal controller of the approximate system will stabilize and yield a suboptimal performance for the original plant. See Wu and Lee (1988) for all the details about this method and the difficulties associated with it.

The techniques and results of this paper are valid for a large class of stable distributed plants. However, when we demonstrate our method in detail with an example, we will specialize to delay systems. For such systems the approach of Flamm (1986) is similar to the one given here. The methods given below have already been applied to a flexible beam problem; see Lenz *et al.* (1989).

The rest of the paper is organized as follows. In the next section we summarize the main results of Foias and Tannenbaum (1989), where the main idea is to use the one step extension theory of Adamjan *et al.* (1971) to characterize the suboptimal solutions to the generalized interpolation problem. In Section 3 we exploit this characterization to illustrate the structure of the optimal/suboptimal H^∞ controllers. We apply our procedure to obtain finite dimensional controllers, for distributed stable plants with invertible rational outer parts in Section 4; plants

* Received 21 April 1989; revised 8 October 1989; revised 28 February 1990; received in final form 31 May 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. F. Curtain under the direction of Editor H. Kwakernaak.

† Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881 U.S.A. Author to whom all correspondence should be addressed.

‡ Department of Electrical Engineering, University of Minnesota, Minneapolis, MN 55455, U.S.A. and Department of Electrical Engineering, Technion, Haifa, Israel.

whose transfer functions are continuous on the boundary (e.g. strictly proper plants with pure delays) are considered in Section 5; and a design example is given in Section 6 to illustrate how to deal with this situation. Finally, in the last section we summarize the results of the paper and make some concluding remarks.

Notation. Our notation is standard. All Hardy spaces of this paper are defined on the unit disc in the usual way. In particular, we consider the systems as transfer functions (of the complex variable z i.e. the Z transform variable). For continuous time systems we can think of this as transfer functions (of the Laplace transform variable s), transformed via bilinear transformation

$s = \frac{1+z}{\tau(1-z)}$, $\tau > 0$, that maps the unit disc to the right half plane. Therefore one can think of all the formulae given in the paper, in terms of the variable z , as an expression in terms of s by simply replacing z by the inverse transformation

$$z = \frac{\tau s - 1}{\tau s + 1}$$

2. PRELIMINARY REMARKS

The sensitivity minimization problem is to find an internally stabilizing controller C such that the following optimum performance is achieved

$$\inf_{C \text{ stabilizing}} \|W(1 + PC)^{-1}\|_\infty =: \mu.$$

See Fig. 1 for the closed loop set-up, where P is the plant to be controlled and W is the weight modelling the disturbances. Here we consider stability in the sense of H^∞ , that is L^2 BIBO stability (bounded energy inputs give rise to bounded energy outputs). We assume that the weight W and its inverse W^{-1} are stable, (i.e. $W, W^{-1} \in H^\infty$), and moreover W is taken to be rational. In this paper we are going to deal with stable plants with nonconstant inner parts. In other words the plant is assumed to have an inner/outer factorization $P = mP_o$, where m is nonconstant inner and P_o is outer. We will further assume that on the unit circle ∂D , P_o has finitely many zeros. Then we can transform the sensitivity minimization problem to a Nehari

problem as follows. We first invoke the Youla parametrization for the controller

$$C = Q_c(1 - PQ_c)^{-1}, \quad Q_c \in H^\infty, \quad (1 - PQ_c) \neq 0.$$

This gives us

$$\mu = \inf_{Q \in H^\infty} \|W - PQ\|_\infty.$$

It is obvious that

$$\mu \geq \max \{|W(b_o)| : b_o \text{ is a zero of } P_o \text{ on } \partial D\} =: \mu_o.$$

We will now make the assumption that $\mu > \mu_o$, which brings us (see Francis and Zames, 1984) to

$$\mu = \inf_{Q \in H^\infty} \|W - mQ\|_\infty. \quad (1)$$

Conversely, from (1) by finding Q_{opt} realizing μ , and by inverting W and P_o , we get the optimal controller C_{opt} which internally stabilizes the system and satisfies

$$\|W(1 + PC_{opt})^{-1}\|_\infty = \mu. \quad (2)$$

One important point which should be emphasized is that when the plant outer part P_o has zeros on the boundary it is only approximately invertible as a stable causal transfer function (in the precise sense explained in Section 5), so a proper optimal controller does not exist in such a case. This leads us to the definition of suboptimality. Given a tolerance $\epsilon > 0$, we say that C_ϵ is *suboptimal* (with tolerance ϵ) if it internally stabilizes the system and satisfies the bound

$$\|W(1 + PC_\epsilon)^{-1}\|_\infty \leq \mu + \epsilon =: \rho. \quad (3)$$

Finding an approximate inverse for the outer part, when it is not invertible in H^∞ , is not the only problem in computing the suboptimal H^∞ controllers. A key reason why we are interested in suboptimal controllers is that the optimal controller (even if the inverse of outer part is proper rational) can be infinite dimensional because inner part of the plant may be infinite dimensional. It is this point which will be considered first: what happens when the inner part m of the plant is infinite dimensional. So, in Section 3 the question of approximating the inverse of P_o will be left aside, and it will be assumed that $P_o^{-1} \in H^\infty$. Furthermore when we study finite dimensional suboptimal controllers in Section 4, we will consider plants whose outer parts are rational and invertible, so we will be dealing only with the infinite dimensionality coming from the inner part. Then, in Sections 5 and 6 we consider the case where P_o is only approximately invertible and may not be rational.

At this point we would also like to discuss the method of first approximating the plant P by say

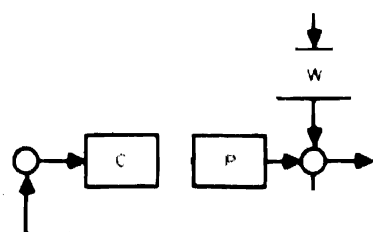


FIG. 1. Closed loop system.

P_f and then checking if the controller obtained from this approximate plant is suboptimal for the original plant. This method requires one to solve

$$\rho = \inf_{Q \in H'} \|W - P_f Q\|_\infty =: \|W - P_f Q_f\|_\infty$$

for ρ and Q_f . On the other hand we know that

$$\begin{aligned} \mu &= \|W - PQ_{\text{opt}}\|_\infty \\ &= \|W - P_f Q_{\text{opt}} - (P - P_f)Q_{\text{opt}}\|_\infty \\ &\leq \|W - P_f Q_{\text{opt}}\|_\infty + \|Q_{\text{opt}}(P - P_f)\|_\infty \\ &\leq \|W - P_f Q_f\|_\infty \\ &\quad + \|P_f(Q_f - Q_{\text{opt}})\|_\infty + \|Q_{\text{opt}}(P - P_f)\|_\infty. \end{aligned}$$

So, if we use the controller $C_f = Q_f(1 - P_f Q_f)^{-1}$ for the plant P the performance ρ will be within ε of μ , provided

$$\varepsilon = \|P_f(Q_f - Q_{\text{opt}})\|_\infty + \|Q_{\text{opt}}(P - P_f)\|_\infty.$$

Similarly, it is easy to see that the stability is preserved if

$$\|Q_f(P - P_f)\|_\infty < 1.$$

The stability condition is well understood even if we do not know the solution to the optimal problem. However, in order to understand the lower bound for ε , given above, we need to find Q_{opt} , that is to solve the optimal problem. This is why in this paper we adopt the method of approximating the optimal controller rather than the method of using the optimal controller obtained by an approximate plant. This way we can compare our finite dimensional controller with the optimal controller, and hence we can observe the effects of approximating the controller on both stability and performance.

In the light of the above discussion, we now assume that P_f is invertible, and consider the following problem: given m inner possibly infinite dimensional, and $\rho < \mu$, find the set of all $Q \in H'$ such that

$$\|W - mQ\|_\infty < \rho. \quad (4a)$$

Let us summarize the results of Foias and Tannenbaum (1989) in connection with the above problem. Suppose that the weight is rational: $W(z) = p(z)/q(z)$ where

$$p(z) = p_0 + zp_1 + \cdots + z^n p_n$$

and similarly

$$q(z) = q_0 + zq_1 + \cdots + z^n q_n$$

(i.e. n is the maximum of the degrees of p and q , so some of the above coefficients may well be zero). Let S denote the unilateral shift on H^2 and define the space $H(m) = H^2 \ominus mH^2$. Then the compressed shift associated with $H(m)$ is

defined as $T := P_{H(m)} S|_{H(m)}$, where $P_{H(m)}$ denotes orthogonal projection.

First consider the optimal case: $\rho = \mu$. The optimal interpolant Q_{opt} , which makes $\|B_{\text{opt}}\|_\infty = \rho$, where

$$B_{\text{opt}} = W - mQ_{\text{opt}},$$

can be computed using Sarason's theorem (Sarason, 1967) which gives that

$$\mu = \|W(T)\|_\infty, \quad W(T) := p(T)q(T)^{-1}.$$

The essential norm is defined as follows:

$$\|W(T)\|_\infty = \sup\{\|W(\zeta)\| : \zeta \text{ singular point of } m\}.$$

We need to assume $\mu > \|W(T)\|_\infty$ (see Foias *et al.*, 1988), to conclude that $W(T)$ attains its norm at a singular value $\rho = \mu$. In this case there exists a singular vector h_ρ , for the so-called (Bercovici *et al.*, 1988) skew Toeplitz operator

$$A_\rho = \rho^2(T)q(T)^* - p(T)p(T)^*$$

(*denotes adjoint) which makes

$$A_\rho h_\rho = 0.$$

The vector h_ρ can be computed explicitly from the problem data $W = p/q$ and m in terms of a determinantal formula; see Foias and Tannenbaum (1989), and Foias *et al.* (1988). Then, B_{opt} can be found via Sarason's result as

$$B_{\text{opt}} = \begin{bmatrix} q(T)^* h_\rho \\ p(T)^* h_\rho \end{bmatrix}.$$

Let us now consider the strictly suboptimal case: $\rho < \mu$. It is obvious that in this case A_ρ is invertible and its inverse can be computed explicitly; again, the formula is given in Foias and Tannenbaum (1989). This is going to be used in the characterization of all the suboptimal solutions $Q_i \in H'$ which make

$$\|W - mQ_i\|_\infty < \rho. \quad (4b)$$

This characterization is obtained using the one step extension procedure of Adamjan *et al.* (1971). Here we want to summarize the method briefly. Set $m_\alpha(z) := zm_\alpha(z)$ and let T_α denote the compression of S to $H(m_\alpha) = H(m) \oplus \mathbb{C}z$. For $\alpha \in \mathbb{C}$ fixed, the problem of finding $B_{\text{opt}}(z, \alpha) = (W - \alpha m - m_\alpha Q_{\text{opt}}^\alpha)(z)$ such that

$$\|B_{\text{opt}}(\cdot, \alpha)\|_\infty = \|(W - \alpha m)(T_\alpha)\| = \rho$$

can be solved using the technique described above for the optimal case. From one step extension theory (Adamjan *et al.*, 1971) we know that the set of all such $\alpha \in \mathbb{C}$ form a circle, say Γ . Furthermore, the equation of Γ can be explicitly calculated. Then, the set of all suboptimal solutions $Q_i \in H'$ satisfying (4) is

obtained in terms of $B_{opt}(z, \phi(u))$:

$$W - mQ_c = B_{opt}(z, \phi(u)),$$

where $\phi(z)$ is a linear fractional map taking the unit circle to Γ , and $u \in H^\infty$, $\|u\|_\infty \leq 1$ is the free parameter. The explicit characterization is as follows. Set

$$g := (\rho^2 q(T) P_{H(m)} q(S)^* - p(T) P_{H(m)} p(S)^*) m,$$

$$g_2 := q_0 p(T) (1 - mm(0)),$$

and

$$h_1 := A_\rho^{-1} g_1, \quad h_2 := A_\rho^{-1} g_2.$$

For a given $\alpha \in \Gamma$ define

$$h_\alpha(z) := m(z) - h_1(z) - \bar{\alpha} h_2(z),$$

and

$$B(z, \alpha) := \frac{\rho^2 q(S)^* h_\alpha}{p(S)^* h_\alpha - \bar{\alpha} q_0}.$$

Then we have the following result.

Proposition 1 (Foias and Tannenbaum, 1989).

The set of all functions of the form

$$B(z) = W(z) - m(z) Q_c(z)$$

with $Q_c \in H^\infty$, such that $\|B\|_\infty \leq \rho$, is given by

$$\left\{ B(z, \alpha) = \frac{\rho^2 q(S)^* h_\alpha}{p(S)^* h_\alpha - \bar{\alpha} q_0}, \quad u \in H^\infty, \|u\|_\infty \leq 1 \right\}$$

where r and η are certain explicitly computable constants. See Foias and Tannenbaum (1989) for the precise formulae. \square

3. STRUCTURE OF THE (SUB)OPTIMAL CONTROLLERS

Using the above parametrization, we are going to obtain the structure of suboptimal controllers. Once more we assume that the inner part of the plant is infinite dimensional, and the outer part is invertible in H^∞ . Then, using the notation of Proposition 1, we set $B_\alpha(z) := B(z, \alpha)$. We can find the controller from the Youla parametrization as $C = Q_c(1 - P Q_c)^{-1}$, where $Q_c \in H^\infty$ is such that

$$B_\alpha = W - P W Q_c.$$

Therefore,

$$C = P^{-1}(B_\alpha^{-1} W - 1).$$

We now study B_α .

$$B_\alpha(z) := \frac{\rho^2 \bar{q}(z) h_\alpha(z) - \rho^2 h_q(z)}{\bar{p}(z) h_\alpha(z) - h_p(z) - q_0 z^n \bar{\alpha}}$$

where $h_q(z)$ and $h_p(z)$ are polynomials of degree $\leq n-1$ and $\bar{q}(z) = z^n q(z^{-1})$ and similarly

$\bar{p}(z) = z^n p(z^{-1})$. Then,

$$P_o C = \frac{1}{m} \left(\frac{\bar{p}(z) h_\alpha(z) - h_p(z) - q_0 z^n \bar{\alpha} p(z)}{\rho^2 \bar{q}(z) h_\alpha(z) - \rho^2 h_q(z) - q(z)} - 1 \right) \cdot \frac{-\lambda(z) h_\alpha(z) - p(z) h_p(z) + \rho^2 q(z) h_q(z) - q_0 z^n p(z) \bar{\alpha}}{(\rho^2 \bar{q}(z) h_\alpha(z) - \rho^2 h_q(z)) q(z)}$$

where $\lambda(z) = \rho^2 \bar{q}(z) q(z) - \bar{p}(z) p(z)$. Recall that $h_\alpha(z) = m(z) - h_1(z) - \bar{\alpha} h_2(z)$. It is easy to see from the inversion of the skew Toeplitz operator A_ρ , that h_1 and h_2 have the following form [see e.g. Lemma 2.1 and Corollary 2.5 of Foias and Tannenbaum (1989)]

$$h_1(z) = \frac{f_1(z) + m(z) F_1(z)}{\lambda(z)}$$

and

$$h_2(z) = \frac{f_2(z) + m(z) F_2(z)}{\lambda(z)}$$

for some f_1, F_1, f_2, F_2 polynomials of degree $\leq 2n$. This gives us

$$P_o C = \frac{1}{m} \left(\frac{-\lambda(z) m(z) (\lambda(z) - F_\alpha(z)) + \lambda(z) \times (-p(z) h_p(z) + q(z) h_q(z) + f_\alpha(z) - \bar{\alpha} q_0 z^n p(z))}{\rho^2 \bar{q}(z) q(z) m(z) (\lambda(z) - F_\alpha(z)) - \rho^2 (\bar{q}(z) q(z) f_\alpha(z) + q(z) h_q(z) \lambda(z))} \right),$$

where

$$F_\alpha(z) := F_1(z) + \bar{\alpha} F_2(z),$$

and

$$f_\alpha(z) := f_1(z) + \bar{\alpha} f_2(z).$$

Since C cannot have poles at the zeros of m (by internal stability) we must have

$$f_\alpha(z) - p(z) h_p(z) + q(z) h_q(z) - \bar{\alpha} q_0 z^n p(z) = 0.$$

It is routine to check that this is satisfied by the interpolation conditions posed in the definition of h_1 and h_2 .

Hence we obtain the following expression

$$P_o C = \left(\frac{-\lambda(z)}{\rho^2 q(z) \bar{q}(z)} \right) \frac{G_\alpha(z)}{1 + m(z) G_\alpha(z)}$$

where

$$G_\alpha(z) = \bar{q}(z) \frac{F_\alpha(z) - \lambda(z)}{\bar{q}(z) f_\alpha(z) + h_q(z) \lambda(z)}$$

and

$$\bar{\alpha} = -\frac{\eta}{u}.$$

Note that

$$\begin{aligned} \frac{-\lambda(z)}{\rho^2 q(z) \tilde{q}(z)} &= \frac{p(z) \tilde{p}(z)}{\rho^2 q(z) \tilde{q}(z)} - 1 \\ &= \frac{W(z) W(z^{-1})}{\rho^2} - 1. \end{aligned}$$

These formulae lead us to the following result.

Proposition 2. Assume that the plant is in the form $P = mP_o$, with m nonconstant inner, and $P_o, P_o^{-1} \in H^\infty$. Then the set of all controllers which internally stabilize the plant P , and satisfy the bound

$$\|W(1 + PC)^{-1}\|_\infty \leq \rho$$

for $\rho \geq \mu$, have the form

$$C = \left(\frac{W(z)W(z^{-1})}{\rho^2} - 1 \right) \frac{G_u(z)}{1 + m(z)G_u(z)} P_o^{-1}(z) \quad (5a)$$

$u \in H^\infty$, $\|u\|_\infty \leq 1$, where $G_u(z)$ is a linear fraction transformation in the free parameter u :

$$G_u(z) = \frac{q_1(z) + q_2(z)u}{q_3(z) + q_4(z)u}$$

with q_1, \dots, q_4 polynomials of degree $\leq 3n$. They can be computed explicitly from the equations given in Foias and Tannenbaum (1989) via f_1, F_1, f_2, F_2, r and η .

Proof. Recall that m is a nonconstant inner function, and $P_o^{-1} \in H^\infty$. From the Youla parametrization, the controller must have the form $C = Q_o(1 - PQ_o)^{-1}$, $Q_o \in H^\infty$. Then the weighted sensitivity function becomes $W(1 - PQ_o)$. Defining $Q_c = WP_oQ_o$, we see that

$$\|W(1 - PQ_o)\|_\infty = \|W - mQ_c\|_\infty.$$

We can use the results of Proposition 1 to characterize all $Q_c \in H^\infty$ which make $\|W - mQ_c\|_\infty \leq \rho$. This in turn characterizes all $Q_o \in H^\infty$ which make $\|W(1 - PQ_o)\|_\infty \leq \rho$. Then the computations in this section give us the structure (5a) for all the controllers $C = Q_o(1 - PQ_o)^{-1}$, $Q_o \in H^\infty$, which satisfy $\|W(1 + PC)^{-1}\|_\infty \leq \rho$. \square

Figure 2 shows the block diagram of the closed loop system with a suboptimal controller. From the structure of the controller we see that if P_o^{-1} is rational and the free parameter u is chosen as a finite dimensional transfer function, then the only infinite dimensional part of the controller is m and it appears at the feedback path around G_u . We thus identify the finite and infinite dimensional parts of the controller. Note that in order to have a finite dimensional controller we

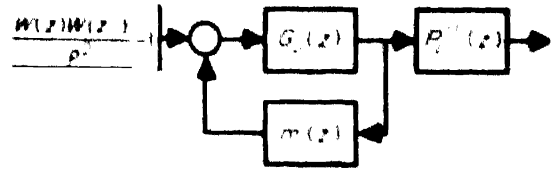


FIG. 2. Structure of the suboptimal controller

must choose the free parameter u depending on m , and moreover u itself must be infinite dimensional, so that the infinite dimensional part of the controller gets cancelled.

As it can be seen from the above formula for C , it is certainly not obvious to characterize the set of all $u \in H^\infty$ and $\|u\|_\infty \leq 1$, such that the transfer function

$$\frac{G_u}{1 + mG_u}$$

is rational. Therefore, we are going to follow a different approach to obtain finite dimensional suboptimal H^∞ controllers. This is the subject of the next section.

4. FINITE-DIMENSIONAL SUBOPTIMAL CONTROLLERS

We now use the above structure for the optimal controller and try to obtain a finite dimensional suboptimal controller. As in the previous section we are going to deal with an infinite dimensional m . Furthermore we will assume that the outer part is rational and invertible. Recall from the Proposition 1 that since $P_o^{-1} \in H^\infty$, the structure described by (5a) is valid for the optimal controller as well. In the case of the optimal controller, however, the free parameter is absent in the term G_u ; that is, instead of G_u we have a fixed function, say G . Hence, since P_o is assumed to be rational, if we replace m by a rational function m_r in the feedback path around G we obtain a finite dimensional controller. We now study the effects of this approximation of m by m_r . More specifically we want to answer the following questions: under which conditions the stability is preserved, and what is the deviation from the optimal performance?

For simplicity of the notation and the computations, we will restrict ourselves to the first order polynomial weights:

$$W(z) = \frac{p(z)}{q(z)}, \quad q(z) = 1, \quad p(z) = p_0 + p_1 z,$$

and $m(z)$ is any arbitrary inner function. The general case is similar, but explicit computation of G requires to solve $2n$ linear equations, where n is the order of the weight (see Section 2), which complicates the notation and the computation considerably. For the above weight W

the optimal sensitivity and the optimal controller are given by the following.

Proposition 3. For a plant $P(z) = m(z)P_o(z)$ where P_o is rational and invertible in H^+ , m arbitrary inner, and the weight $W(z) = p(z) = p_o + p_1 z$ the optimal sensitivity B_{opt} is

$$B_{opt}(z) = W(z) \frac{1 + m(z)\mu z/p(z)}{1 + m(z)\bar{p}(z)/\mu}.$$

Consequently the optimal controller is

$$C_{opt}(z) = \left(\frac{W(z)W(z^{-1})}{\mu^2} - 1 \right) \times \frac{G(z)}{1 + m(z)G(z)} P_o^{-1}(z) \quad (5b)$$

with

$$G(z) := \frac{\mu z}{p(z)}.$$

The optimal performance μ can also be computed explicitly.

Proof. See Appendix A. \square

Now we replace m by m_f in the expression for the controller, so that the controller becomes a finite dimensional transfer function:

$$C_f(z) := \left(\frac{W(z)W(z^{-1})}{\mu^2} - 1 \right) \times \frac{\mu z/p(z)}{1 + m_f(z)\mu z/p(z)} P_o^{-1}(z).$$

It is easy to see that if we use C_f in the closed loop as a controller, then the sensitivity function becomes

$$B_f(z) = W(z) \left(\frac{1 + m(z)\mu z/p(z) + \Delta}{1 + m(z)\bar{p}(z)/\mu + \Delta} \right)$$

where $\Delta = (m_f(z) - m(z))\mu z/p(z)$.

Set

$$R(z) := \frac{p(z)}{\mu} \left(1 + m(z) \frac{\bar{p}(z)}{\mu} \right),$$

$$\Delta_m(z) := z(m_f(z) - m(z)),$$

and

$$X(z) := \frac{\Delta_m(z)}{R(z)}.$$

Then we can rewrite B_f as

$$B_f(z) = B_{opt}(z) \frac{1}{1 + X(z)} + W(z) \frac{X(z)}{1 + X(z)}.$$

This expression shows that for suboptimality of C_f , the rational function m_f should be designed by studying the relation between the terms R and Δ_m .

From this point on, in the examples that we are going to consider, we will conduct our analysis and design in the right half plane, which is more natural for continuous time systems. When we do this we transform the problem data

by using the conformal map $z = \frac{ts-1}{ts+1}$ between

the right half plane and the unit circle. In particular $\hat{R}(s)$ denotes $R(z) \big|_{z=(s-1)/(s+1)}$, and $\Delta_m(s) := \Delta_m(z) \big|_{z=(s-1)/(s+1)}$, and similarly for all the other transfer functions. Let us now compare \hat{R} and Δ_m to analyze the suboptimality of \hat{C}_f .

First of all in order to guarantee stability we should have

$$1 + \hat{X}(s) \neq 0$$

inside the closed right half plane, a sufficient condition would be $\hat{X} \in H^+$, $\|\hat{X}\|_\infty < 1$.

Also, since we are looking for a suboptimal controller, H^+ norm of \hat{B}_f should be close to μ . Note that if we could make $1 \gg \|\hat{X}\|_\infty$, i.e.

$$|\hat{R}(j\omega)| \gg |\Delta_m(j\omega)|$$

for all $\omega \geq 0$, then we would have

$$|\hat{B}_f(j\omega)| \approx |\hat{B}_{opt}(j\omega)| \quad \forall \omega \geq 0$$

which implies that $\|\hat{B}_f\|_\infty \approx \|\hat{B}_{opt}\|_\infty = \mu$. However this is not possible in general because there is no good uniform (on the imaginary axis) rational approximation for an irrational inner function which has essential singularities on the boundary. This is the main difficulty in finding the finite dimensional suboptimal H^+ controllers for distributed systems with invertible outer parts. To illustrate this point we would like to give the following example.

Example. Choose the weight to be the first order low-pass filter

$$\hat{W}(s) = \frac{\varepsilon_\omega \tau_\omega s + 1}{\tau_\omega s + 1}$$

where $0 < \varepsilon_\omega < 1$, $\tau_\omega > 0$, and take the plant to be

$$\hat{P}(s) = e^{-hs} P_0(s)$$

with P_0 rational and invertible in H^+ . We use $z = \frac{\tau_\omega s - 1}{\tau_\omega s + 1}$ in going from right half plane to the disc. Therefore, we compare

$$\hat{R}(s) = \frac{1}{\mu} \left(\frac{\varepsilon_\omega \tau_\omega s + 1}{\tau_\omega s + 1} \right) \left(1 + \frac{e^{-hs} \varepsilon_\omega \tau_\omega s - 1}{\mu (\tau_\omega s + 1)} \right)$$

and

$$\Delta_m(s) = \frac{\tau_\omega s - 1}{\tau_\omega s + 1} (e^{-hs} - \hat{m}_f(s))$$

where $\hat{m}_r(s)$ is a finite dimensional (e.g. Padé) approximation for the delay term $e^{-\mu s}$. We observe that $|\hat{R}(j\omega)|$ oscillates around $\epsilon_\infty/\tau_\infty < 1$ with oscillation amplitudes $\epsilon_\infty^2/\tau_\infty^2$ as $\omega \rightarrow \infty$. On the other hand, assuming \hat{m}_r is also inner, $|\Delta_m(j\omega)|$ oscillates between 0 and 2 as $\omega \rightarrow \infty$. So, we cannot guarantee stability nor good performance because of the essential singularity of $\hat{m}(s) = e^{-\mu s}$ at $s = j\infty$ [or $\hat{m}(z) = e^{\mu \ln(z) + 1/(z-1)}$ has essential singularity at $z = e^{j0} = 1$ on the unit circle]. It is worth mentioning once more that in this example the problem arises in the high frequency range, i.e. near the point where the plant has a discontinuity. In the next section, we generalize the above idea, of designing \hat{m}_r by comparing \hat{R} with Δ_m , to plants which can be approximated uniformly on the imaginary axis. In this case although the inner part of the plant may have essential singularities, the outer part is such that the plant itself does not have any discontinuity on the boundary, and we will see that in this manner we can solve our problem.

5. ON THE OUTER PART OF THE PLANT

In light of the above discussion we now consider the case in which the plant \hat{P} is continuous on the imaginary axis. In other words the outer part of the plant is such that the essential singularities of the inner part get killed. The assumption that \hat{P} is continuous guarantees that \hat{P}_o is also continuous (Garnett, 1981, p. 78), and hence both \hat{P} and \hat{P}_o are uniformly approximable by rational functions on the imaginary axis (Hoffman, 1988, p. 77), i.e. for any given ϵ there exists $\hat{P}_r \in H^\infty$ rational such that $\|\hat{P} - \hat{P}_r\|_\infty < \epsilon$.

In this section, in order to simplify our exposition we will consider strictly proper (continuous time) plants, with no zeros on the finite imaginary axis, and with inner parts having only one essential singularity which is at $|s| = \infty$, $\text{Re}(s) \leq 0$. A typical example is a plant with transportation delay, and strictly proper continuous outer part. However, the discussion below can be easily extended to continuous plants having finitely many zeros on the boundary and in which the inner part may have essential singularities at these points.

As before we work with data transformed from the right half plane into the unit disc, in such a way that the point $|s| = \infty$, $\text{Re}(s) \leq 0$ is mapped to $z = e^{j0} = 1$.

Recall that the inverse of the outer part appears in the controller expression. Note that P_o being outer cannot have zeros in the interior of the unit disc. But it may have zeros on the boundary and in this case it is not invertible in

H^∞ . Following our above remarks we assume that P_o has one zero on ∂D (at which the inner part of the plant may have an essential singularity.) Once again the argument which we will now give works for a finite number of zeros (at which the inner part is allowed to have essential singularities).

We will now employ an "approximate" inverse, in the following sense. Let N be a small open neighborhood of the point on ∂D where P_o has a zero. Then since P_o is outer and continuous on the boundary, its inverse P_o^{-1} exists in $D_p := D \setminus N$ as a bounded analytic function which is moreover continuous on \bar{D}_p . Now note that

$$\sup_{z \in \bar{D}_p} |P_o(z)| \approx \sup_{z \in \partial D} |P_o(z)| = \|P_o\|_\infty < \infty.$$

Then from a corollary to Wermer's maximality theorem (Hoffman, 1988), P_o^{-1} is a uniform limit of polynomials on \bar{D}_p . Thus given any such small neighborhood N , and any $\epsilon > 0$ we can find a polynomial function, which we denote by P_{of}^{-1} , such that

$$|P_o(z)P_{of}^{-1}(z) - 1| \leq |P_{of}^{-1}(z) - P_o^{-1}(z)| |P_o(z)| \leq \epsilon \|P_o\|_\infty$$

uniformly in $z \in \bar{D}_p$. Note that once N is fixed we can make the left hand side of the above inequality arbitrarily close to zero by choosing ϵ sufficiently small. On the other hand, for $z \in N$, $P_o(z)P_{of}^{-1}(z)$ need not be close to 1. In fact as $z \in N$ approaches the point where P_o has the zero, we have that $P_o(z)P_{of}^{-1}(z)$ approaches zero, because of continuity and the fact that P_{of}^{-1} is a polynomial. In summary, such a function P_{of}^{-1} will be called an *approximate inverse* for the outer part P_o of the given plant. This approximation is uniform on the boundary excluding any arbitrarily small neighborhood around the point on ∂D where P_o has the zero. In case P_o has a finite number of zeros we take N to consist of a finite union of small open neighborhoods about each zero. The above construction then of the approximate inverse goes through exactly as just described.

Now returning to the problem of approximating the optimal controller, recall that structure of the optimal controller is given by:

$$C(z) = \left(\frac{W(z)W(z^{-1})}{\mu^2} - 1 \right) \frac{G(z)}{1 + m(z)G(z)} P_o^{-1}(z).$$

We can rewrite this as

$$C(z) = \left(\frac{W(z)W(z^{-1})}{\mu^2} - 1 \right) \frac{G(z)P_o^{-1}(z)}{1 + P(z)G(z)P_o^{-1}(z)}.$$

A proper finite dimensional controller can be

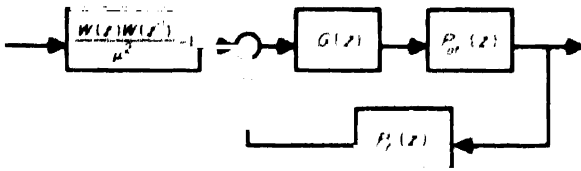


FIG. 3 Finite dimensional controller.

obtained by approximating P and P_o^{-1} separately:

$$C_f(z) = \left(\frac{W(z)W(z^{-1})}{\mu^2} - 1 \right) \frac{G(z)P_o^{-1}(z)}{1 + P_f(z)G(z)P_o^{-1}(z)},$$

where P_f and P_o^{-1} are finite dimensional proper approximations for P and P_o^{-1} respectively. See Fig. 3 for the implementation of C_f . Of course, in order to physically implement C_f , we will have to assume from now on that the rational function G is proper. This assumption automatically holds for the first order polynomial weights case studied in Section 4. Indeed, there $G(z)$ was proper because the inverse of the weight was proper. In fact, we believe that the rational function G appearing in the above formula for the structure of the optimal controller is proper whenever the inverse of the weight is proper. However, at this point we have no rigorous proof of this.

Let us now analyze the sensitivity function for the closed loop system under this finite dimensional controller. After a simple algebraic computation similar to the one given in Section 4, we see that the sensitivity function $B_f := W(1 + PC_f)^{-1}$ is in the form

$$B_f(z) = B_{opt}(z) \frac{1}{1 + X_1(z)} + W(z) \frac{X_2(z)}{1 + X_1(z)}, \quad (6)$$

where $X_1 = \Delta/R$ and $X_2 = \delta/R$ with

$$R(z) := \left(1 + \frac{p(z)\tilde{p}(z)}{\mu^2 q(z)\tilde{q}(z)} m(z)G(z) \right) G(z)^{-1},$$

$$\delta(z) := m(z)\delta_{of}(z) + \delta_p(z)P_o^{-1}(z),$$

$$\Delta(z) := \frac{p(z)\tilde{p}(z)}{\mu^2 q(z)\tilde{q}(z)} m(z)\delta_{of}(z) + \delta_p(z)P_o^{-1}(z),$$

$$\delta_{of}(z) := P_o(z)P_o^{-1}(z) - 1,$$

$$\delta_p(z) := P_f(z) - P(z).$$

Following the above ideas, in order to make C_f suboptimal we design P_f and P_o^{-1} by comparing R with Δ and δ . A sufficient condition for stability is $X_1, X_2 \in H^\infty$ and $\|X_1\|_\infty < 1$.

The analysis for the performance is similar to the case studied in Section 4 with one important difference. Since $P(e^{j\phi})$ is continuous on $\phi \in [-\pi, \pi]$, we can approximate it up to any given tolerance by a finite dimensional transfer function P_f , uniformly on the unit circle ∂D . Similarly, from the discussions on the approxi-

ate inverse, $\delta_{of}(e^{j\phi})$ can be made arbitrarily close to zero away from the point $z = e^{j\theta}$ where P_o has a zero. On the other hand, because of this zero, and since P_o^{-1} is in H^∞ , the controller C_f is proper and the loop transfer function PC_f is strictly proper. Therefore

$$\|B_f(e^{j\phi}) - W(e^{j\phi})\| \rightarrow 0 \quad \text{as } \phi \rightarrow 0.$$

We also know that when $P_o(e^{j\theta}) = 0$, by assumption $\mu > \mu_o$, we necessarily have $|W(e^{j\theta})| < \mu$. Therefore, in this case having a proper $P_o^{-1} \in H^\infty$ guarantees a good performance near the point where inner part has an essential singularity and outer part has a zero. Also away from this point, since we have "good" approximations for P and P_o^{-1} , equation (6) shows that we can obtain a level of performance close to the optimum. Again from (6) we see that in the transition region on the unit circle where $P_o P_o^{-1}$ is neither close to 1 nor to 0 a deviation from the optimum performance may occur. It is possible to obtain at least a conservative bound from (6) immediately:

$$\|B_f\|_\infty \leq (\|B_{opt}\|_\infty + \|WX_2\|_\infty) \left(\frac{1}{1 - \|X_1\|_\infty} \right).$$

However this bound is by no means tight. Better bounds on this performance degradation should be obtained by studying on what happens to the magnitudes of the approximation $P - P_f$, the weight W and G in this frequency band.

In the next section we illustrate the method discussed above by considering an example similar to the one studied in Section 4.

6. EXAMPLE: LOW PASS WEIGHTS AND DELAY SYSTEMS

In this section we will consider a first order low pass weight and a plant with delay. We take the outer part of the plant to be strictly proper, so that the transfer function $\hat{P}(s)$ becomes continuous on the imaginary axis.

Let us choose the weight to be

$$\hat{W}(s) = \frac{\epsilon_w \tau_w s + 1}{\tau_w s + 1}$$

and look at the plant

$$\hat{P}(s) = \frac{1}{\tau_p s + 1} \frac{s - b}{s + b}$$

Here h is the amount of time delay, $1/\tau_p$ is the bandwidth of the plant, and $1/\tau_w$ is the bandwidth of the weight (determining the band on which disturbance signals act). Typically ϵ_w is much less than 1. Note that if $\tau_p \neq 1/b$ then we can write $\hat{P}(s) = Ae^{-hs}/(\tau_p s + 1) + Be^{-hs}/(s + b)$

where $A = \frac{1 + \tau_p b}{1 - \tau_p b}$ and $B = \frac{-b}{1 - \tau_p b}$. Therefore the Hankel operator associated with this plant is compact, because the corresponding impulse response is in $L_1(0, \infty)$, (see Adamjan *et al.*, 1971).

For this example the conformal map between unit circle and right half plane will be taken as

$$z = \frac{\tau_w s - 1}{\tau_w s + 1} \quad \text{and} \quad s = \frac{1}{\tau_w} \frac{1 + z}{1 - z}.$$

This puts the weight in the form $W(z) = p(z) = p_0 + p_1 z$, with $p_0 = (1 + \epsilon_w)/2$ and $p_1 = -(1 - \epsilon_w)/2$. In this case we have $G(z) = \mu z/p(z)$ (see Appendix A). The method we are using here does not depend on the conformal map chosen. However the above choice for the conformal map makes $W(z)$ a polynomial, and simplifies the computations of the Appendix A.

A natural choice for $\hat{P}_f(s)$ is $\hat{m}_f(s)/(\tau_p s + 1)$, where \hat{m}_f is a finite dimensional approximation of the inner part of the plant. As an approximate inverse of the outer part we choose the proper function

$$\hat{P}_{ot}^{-1}(s) = \frac{\tau_p s + 1}{\epsilon_p \tau_p s + 1}$$

where $\epsilon_p > 0$ is very small (we discuss later how small this should be).

Now we check under which conditions the finite dimensional controller

$$C_f(z) = \left(\frac{p(z)\bar{p}(z)}{\mu^2 z} - 1 \right) \frac{G(z)P_{ot}^{-1}(z)}{1 + G(z)P_f(z)P_{ot}^{-1}(z)}$$

is suboptimal. Recall the equation (6), from which we have

$$B_f = B_{opt} \frac{1}{1 + X_1} + \frac{X_2}{1 + X_1}$$

where $X_1 := \Delta/R$, and $X_2 = W\delta/R$. Therefore, if the conditions

- (a) $X_1 \in H^\infty$
- (b) $X_2 \in H^\infty$
- (c) $\|X_1\|_\infty < 1$

are satisfied, then $B_f \in H^\infty$. Assuming $m_f \in H^\infty$, then for (a) and (b) to hold it is necessary and sufficient to have

$$\hat{m}_f(j\omega_c) = \hat{m}(j\omega_c)(1 + j\epsilon_p \tau_p \omega_c) \quad (7)$$

(see Appendix B), where ω_c is determined by the zeros of $(\bar{p}(z)p(z) - \mu^2 z)$ for $z = (\tau_w s - 1)/(\tau_w s + 1)$. Simple computations give that

$$\omega_c = \frac{y}{\tau_w} \quad \text{with} \quad y = \sqrt{\frac{1 - \mu^2}{\mu^2 - \epsilon_w^2}}.$$

We will choose a Padé approximation for the delay term and add a filter to this to take into account the effect of $(1 + j\epsilon_p \tau_p \omega_c)$:

$$\hat{m}_f(s) = \frac{s - b}{s + b} \hat{m}_d(s) \hat{F}(s)$$

where

$$\hat{F}(s) = \frac{s^2 + 2\omega_c s + (\omega_c^2 - r_i^2)}{s^2 + 2\omega_c s + \omega_c^2}$$

and \hat{m}_d is a Padé approximation which is going to be defined below. The choice of $r_i = \omega_c \sqrt{2\epsilon_p \tau_p \omega_c}$ makes $\hat{F}(j\omega_c) = (1 + j\epsilon_p \tau_p \omega_c)$. So, we need only to check if, say the first order, Padé approximation $(1 - jh\omega_c/2)/(1 + jh\omega_c/2)$ is actually equal to $e^{jh\omega_c}$. This does not in general hold, however when h is order of magnitude 0.01 (and ω_c is less than 0.1) then the difference is so small (less than 10^{-10}) that we can fix the problem by changing the term $(1 - hs/2)/(1 + hs/2)$ to

$$K \frac{1 - \left(\frac{h}{2} + d_i\right)s}{1 + \frac{h}{2}s} =: \hat{m}_d(s).$$

Here for such small values of h and ω_c we have $1 - K$ and d_i are less than 10^{-10} . So, $\hat{m}_d(j\omega_c)$ can practically be seen to be equal to the first order Padé approximation of $e^{jh\omega_c}$. In summary, we are going to use

$$\hat{m}_f(s) = \frac{s - b}{s + b} \hat{m}_d(s) \hat{F}(s)$$

in the controller, where \hat{m}_d defined above is a first order approximation for the delay term. Note that we need the exact values of K and d_i to satisfy the condition (a) and (b).

One other condition we need to satisfy is $\|X_1\|_\infty < 1$. After substitution of the terms we see that

$$\hat{X}_1(s) = \frac{\left(\frac{1 - \epsilon_w \tau_w s - 1}{\mu - \tau_w s + 1} \frac{-\epsilon_p \tau_p s}{\epsilon_p \tau_p s + 1} \hat{m}(s) + \mu \frac{\tau_w s - 1}{\epsilon_w \tau_w s + 1} \frac{\hat{m}_f(s) - \hat{m}(s)}{\epsilon_p \tau_p s + 1} \right)}{1 + \frac{1 - \epsilon_w \tau_w s - 1}{\mu - \tau_w s + 1} \hat{m}(s)}$$

Define

$$\hat{R}_c(s) := 1 + \frac{1 - \epsilon_w \tau_w s - 1}{\mu - \tau_w s + 1} \hat{m}(s),$$

and

$$\hat{D}_c(s) := \hat{X}_1(s) \hat{R}_c(s).$$

Plot $|\hat{R}_c(j\omega)|$ versus ω , and choose $\epsilon_p \tau_p$ small enough such that

$$|\hat{D}_c(j\omega)| \ll |\hat{R}_c(j\omega)| \quad \text{for} \quad \omega \leq \omega_c$$

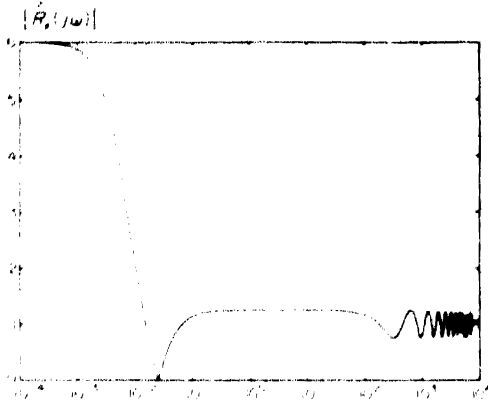


FIG. 4. $|R_e(j\omega)|$ versus ω .

ω_r is to be defined below. Consider the second term in the numerator of \hat{X}_1 . Note that

$$\mu \left| \frac{\tau_w j\omega - 1}{\epsilon_w \tau_w j\omega + 1} \right| \rightarrow \frac{\mu}{\epsilon_w} \quad \text{as } \omega \rightarrow \infty.$$

So, \hat{m}_r should approximate \hat{m} "reasonably well" at least up to frequency ω_r where

$$\epsilon_p \tau_p \omega_r \gg \frac{2\mu}{\epsilon_w}.$$

Then, $|\hat{D}_e(j\omega)| \approx \epsilon_w/\mu$ (which is going to be strictly smaller than $1 - \epsilon_w/\mu = |R_e(j\omega)|$ when $\mu > 2\epsilon_w$) for $\omega > \omega_r$. We therefore assume that $\mu > 2\epsilon_w$. All these guarantee only the stability of B_r .

At this point we should also check suboptimality, i.e. consider $|\hat{B}_r(j\omega)|$. Recall the expression

$$\hat{B}_r = \hat{B}_{opt} \frac{1}{1 + \hat{X}_1} + \frac{\hat{X}_2}{1 + \hat{X}_1}$$

where \hat{X}_1 is as above and \hat{X}_2 can be computed as

$$\hat{X}_2(s) = \mu \frac{\tau_w s - 1}{\tau_w s + 1} \frac{\frac{\hat{m}(s) - \epsilon_p \tau_p s}{\epsilon_p \tau_p s + 1} + \frac{\hat{m}_r - \hat{m}}{\epsilon_p \tau_p s + 1}}{1 + \frac{1 - \epsilon_w \tau_w s - 1}{\mu - \tau_w s + 1} \hat{m}(s)}.$$

Now, since at low frequencies $|\hat{X}_1(j\omega)| \ll 1$ and $|\hat{X}_2(j\omega)| \ll 1$, by sufficiently small ϵ_p and good approximation of \hat{m} by \hat{m}_r , we have

$$|\hat{B}_r(j\omega)| \approx |\hat{B}_{opt}(j\omega)|$$

for low frequencies. It is also not difficult to see that at high frequencies

$$|\hat{B}_r(j\omega)| \rightarrow \epsilon_w < \mu$$

because $\left| \frac{1}{j\epsilon_p \tau_p \omega + 1} \right| \rightarrow 0$.

Thus, recapping, choosing ϵ_p sufficiently small (for good performance at low frequencies) and approximating $\hat{m}(j\omega)$ by $\hat{m}_r(j\omega)$ up to the

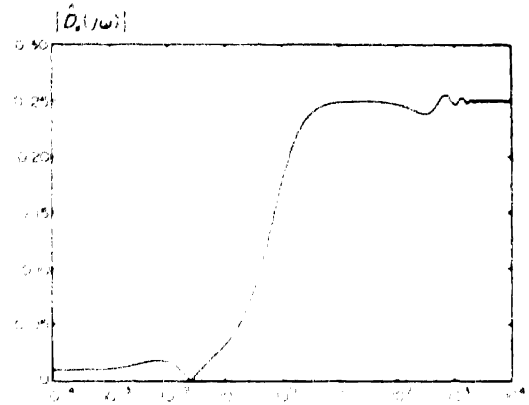


FIG. 5. $|D_e(j\omega)|$ versus ω .

frequency range near $\omega_r \gg \frac{2\mu}{\epsilon_w \epsilon_p \tau_p}$, also satisfying the interpolation condition (7) posed by stability, we have suboptimality.

We remark about a trade-off now. For good performance we need to use small ϵ_p ; however, this increases the frequency band on which \hat{m} should be approximated well, which forces us to use higher order approximations if the delay h is not small enough.

Let us look at a specific design example by choosing $\tau_w = 200$, $\tau_p = 100$, $\epsilon_w = 0.05$, $h = 0.01$, $\epsilon_p = 0.01$, and $b = 0.0267$. These make $\omega_r = 1/39.5$ and $\mu = 0.2$ (see Appendix A). The values for d_i and K can be computed by equating the magnitude and phase of $e^{j\omega h}$ to the magnitude and phase of $\hat{m}_d(j\omega_r)$. We find that $K = (1 + 5.530 \times 10^{-11})^{-1}$ and $d_i = 2.765 \times 10^{-11}$. After substitution of all the terms in (6) and simplifications the above method leads to the controller given by

$$\hat{C}_f(s) = -7.5 \left(\frac{(200 + s)(100s + 2.67) \times (100s + 1)(s + 0.025)^2}{(191.8 + s)(200s + 1)(s + 5.35) \times (s + 0.03)(s + 0.02)} \right).$$

The magnitude plots of \hat{R}_e , \hat{D}_e and \hat{X}_1 are given in Figs 4, 5, 6 respectively; we observe the

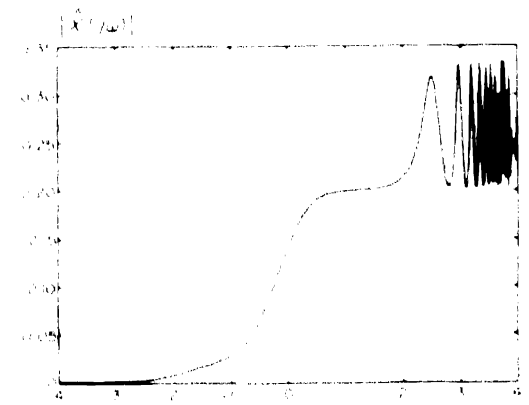


FIG. 6. $|\hat{X}_1(j\omega)|$ versus $\log(\omega)$.

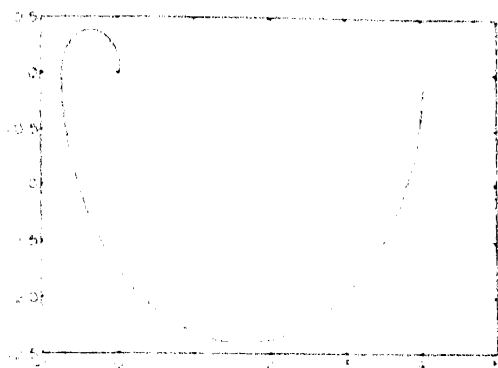


Fig. 7. Nyquist plot for $\tilde{P}\tilde{C}$.

stability from these plots (note also that there is no unstable pole-zero cancellation in the controller and the plant). We can also check stability from the Nyquist plot for this controller and plant, Fig. 7. For the performance bound see the magnitude plot of \tilde{B}_t in Fig. 8. We have $\|\tilde{B}_t\|_\infty = 0.208$, so the deviation from the optimal performance is about 4%. If better performance is desired (i.e. $\rho < 0.208$) then one should decrease the value of ϵ_p and refine the approximation of the delay term accordingly (if necessary).

Remark. For plants with invertible outer parts we have seen in Section 4 that there is a difficulty in our method. Nevertheless, we can overcome this difficulty in the following way. Assume that the outer part P_o of the plant P is invertible and continuous on the boundary, and furthermore suppose that the inner part m has finitely many essential singularities on the unit circle ∂D . Then, there exists a rational outer transfer function P_r such that $P_r m$, and hence $P_r P$ are continuous on ∂D . One can construct this function by simply defining $P_r = \prod (z - \zeta_i)$, where ζ_i s are precisely the essential singularities of m on ∂D . Since P_r is outer the optimal performance $\mu(P)$ corresponding to the plant P is the same as the optimal performance $\mu(P_r P)$.

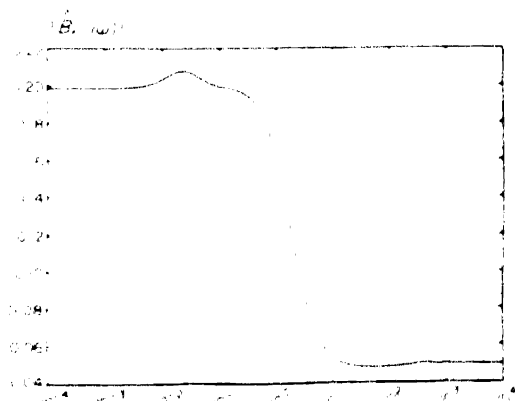


Fig. 8. $\|\tilde{B}_t(j\omega)\|$ versus ω .

for the plant $P_r P$. (This is true by the assumption that $\mu > \|W(T)\|_\infty$, and P_r can be taken in such a way that it vanishes only at those points where m has essential singularities on the boundary.) Moreover, the corresponding optimal H^∞ controllers are related as follows:

$$C_{\text{opt}}(P) = C_{\text{opt}}(P_r P) P_r^{-1}$$

A finite dimensional suboptimal controller for the plant P can be found by approximating $C_{\text{opt}}(P_r P)$:

$$C_f(P) = C_f(P_r P) P_r^{-1}$$

where the structure of $C_f(P_r P)$ is as in Fig. 3, with $P_r P_o$ taking place of P_o . In Sections 5 and 6 of this paper we discussed the problem corresponding to this situation where the plant is continuous, e.g. $P_r P$, and the controller is in the form of $C_f(P_r P)$. So, in summary, for plants with continuous invertible outer parts a suboptimal finite dimensional controller can be found by introducing a rational outer function P_r , which makes $P_r P$ continuous on the unit circle, and by treating $P_r P_o$ as the outer part.

7. CONCLUDING REMARKS

In this paper we have considered the one block H^∞ sensitivity minimization problem in the SISO case. We have obtained the structure of all the suboptimal H^∞ controllers for systems with rational weights and stable arbitrary distributed plants, having invertible outer parts. From this structure we have identified the infinite and finite dimensional parts of the optimal controller as follows

$$C = \left(\frac{W(z)W(z^{-1})}{\mu^2} - 1 \right) \frac{G(z)}{1 + m(z)G(z)} P_o^{-1}(z).$$

Although we have made the observation that G is finite dimensional and can be obtained from the $2n$ -equations of the singular system, Foias *et al.* (1988), we could not find an explicit formula for $G(z)$ in case of arbitrary weights W , because of the complexity of the equations. So, there is more work to be done in this direction.

In the first part of our discussion on this structure (Section 4) we have assumed a first order weight and proper rational and stable P_o^{-1} . We computed $G(z)$ as $\mu z / W(z)$. Then we have studied under which conditions the finite dimensional controller

$$C_f = \left(\frac{W(z)W(z^{-1})}{\mu^2} - 1 \right) \frac{G(z)}{1 + m_f(z)G(z)} P_o^{-1}(z)$$

is suboptimal where m_f is a rational function approximating m . We have derived that when C_f is used in the closed loop, then the sensitivity

becomes

$$B_f = B_{opt} \frac{1}{1+X} + W \frac{X}{1+X} \quad (8)$$

where $X = \Delta_m/R$ with $\Delta_m = m_f - m$ and $R = G^{-1}(1 + mGW(z)W(z^{-1})/\mu^2)$, and B_{opt} is the optimal sensitivity. The above formula (8) is true even if the weight is higher order rational function, but the expression for $G(z)$ (which is equal to $\mu z/W(z)$ if the weight is a first order polynomial) is more complicated.

One of the main arguments of this paper is that m_f should be designed by comparing R with Δ_m . We have seen that there is a difficulty in the case of plants whose outer parts are invertible. That is we cannot have

$$|\hat{R}(j\omega)| \gg |\Delta_m(j\omega)|$$

uniformly on the imaginary axis if the inner part of the plant m has essential singularities on the boundary. Nevertheless, as we have explained in the Remark at the end of the previous section, we can get around this problem by looking at the finite dimensional controllers for *stable plants with continuous transfer functions*, whose outer parts need not be rational. In that case we need to invert the outer part approximately, because it will have zeros at the points where the inner part has essential singularities. This issue, of approximately inverting the outer part of the plant, have been addressed in Sections 5 and 6. Basically, when the plant transfer function P is continuous on the unit circle, it is possible to write B_f similar to (8) [see equation (6)]. We have seen that having a proper "approximate inverse" for the outer part, which have zeros on the boundary, helps solving the problem we have encountered in the case where outer part was invertible.

As discussed in Sections 5 and 6 at "high" and "low" frequencies the magnitude of the sensitivity function obtained in this method can be made arbitrarily close to (or less than) the optimal performance μ . we were not able to obtain a tight bound on the deviation from the optimal performance in the "transition range" where $|P_o P_o^{-1}|$ rolls-off to zero. In order to obtain such a bound we need to study the behavior of G , W , m , P and the approximating functions P_f and P_o^{-1} in this frequency band.

Bounds on the deviation from the optimal performance depend also on the approximation bounds. On the other hand approximation to infinite dimensional systems is itself of independent interest. At this point we want to give an idea about how large the class of systems considered in Section 5, i.e. stable (in the H^*

sense) plants with continuous transfer functions, and the approximation problem related to these systems. We believe that the main restriction here is the stability. The other condition, continuity of the transfer function, is required for the existence of the uniform (in the H^* sense) approximations. Because if a distributed system has a discontinuity then it is not possible to approximate it uniformly by a finite dimensional system, see a discussion in Gu *et al.* (1989). Moreover, in order to obtain certain specific bounds in the approximation method used, one may need to make more restricting assumptions than continuity. In this paper we did not address these issues [see for example Gu *et al.* (1989) and Partington *et al.* (1988) and their references for detailed discussions]. When we have illustrated our approach by considering an example in Section 6, we have only used the classical Padé approximation. Other types of approximation schemes, might lead to a better performance while keeping the same order. However, the bottom line is that, in order to find suboptimal controllers, no matter which approximation method is used, one should take into account the structure of optimal/suboptimal controllers given above.

Finally, we want to close our discussion with an important open problem related to the structure obtained in Proposition 2: Given $\rho > \mu$ characterize the set of all $u \in H^*$, $\|u\|_1 \leq 1$, such that the transfer function

$$\frac{G_u}{1 + mG_u}$$

in (5a) is finite dimensional, and find the lowest possible dimension.

Acknowledgements—This work was supported by the National Science Foundation under grants No. ECS-8704047, DMS-8811084, by the Air Force Office of Scientific Research under grants AFOSR-88-0020 and AFOSR-90-0024, and by the Army Research Office

REFERENCES

- Adamjan, V. M., D. Z. Arov and M. G. Krein (1971). Analytic properties of Schmidt pairs for a Hankel operator and generalized Shur-Takagi problem. *Math. USSR Sbornik* **15**, 31–73.
- Bercovici, H., C. Foias and A. Tannenbaum (1988). On skew Toeplitz operators I. *Operator Theory, Adv. Applic.*, **32**, 21–43.
- Flamm, D. S. (1986). Control of delay systems for minimax sensitivity. Ph.D. thesis, MIT, 1986.
- Foias, C. and A. Tannenbaum (1989). On the parametrization of the suboptimal solutions in generalized interpolation. *Linear Algebra Applic.*, **122/123/124**, 145–164.
- Foias, C., A. Tannenbaum and G. Zames (1988). Some explicit formulae for the singular values of a certain Hankel operators with factorizable symbol. *SIAM J. Math. Anal.*, **19**, 1081–1091.
- Francis, B. and G. Zames (1984). On H^* optimal sensitivity theory for SISO feedback systems. *IEEE Trans. Aut. Control*, **AC-29**, 9–16.

- Garnett, J. B. (1981). *Bounded Analytic Functions*. Academic Press, New York.
- Gu, G., P. P. Khargonekar and E. B. Lee (1989). Approximation of infinite dimensional systems. *IEEE Trans. Aut. Control*, **AC-34**, 610-618.
- Hoffman, K. (1988). *Banach Spaces of Analytic Functions*. Dover, New York.
- Lenz, K., H. Ozbay, A. Tannenbaum, J. Tun and B. Morton (1989). "Robust control design for a flexible beam using a distributed parameter H^∞ method. Proceedings of the 28th CDC, Tampa, FL, 2673-2678.
- Partington, J. R., K. Glover, H. J. Zwart and R. F. Curtain (1988). L^∞ approximation and nuclearity of delay systems. *Syst. Control Lett.*, **10**, 59-65.
- Sarason, D. (1967). Generalized interpolation in H^∞ . *Trans. AMS* **127**, 179-203.
- Wu, N. E., and E. B. Lee (1988). Feedback minimax synthesis for distributed systems. Proc. 27th CDC, Austin TX, 492-496.

APPENDIX A

Optimal sensitivity for $W(z) = p(z) = p_0 + p_1 z$

We have seen that computation of the optimal sensitivity reduces to finding a nonzero vector h_0 such that $A_0 h_0 = 0$. That is, in our case,

$$((p_0 + p_1 T)(p_0 + p_1 T^*) - \mu^2 I)h_0 = 0 \quad (a1)$$

But we also have the following

$$Th_0 = zh_0(z) - m(z)u_{-1} \quad (a2 a)$$

$$T^*h_0 = z^{-1}h_0(z) - h_0(0) \quad (a2 b)$$

and

$$TT^*h_0 = h_0(z) - h_0(0)(1 - m(z)m(0)) \quad (a2 c)$$

for some constants u_{-1} and $u_1 = h_0(0)$. Putting these expressions in (a1) we see that (a1) is equivalent to

$$\begin{aligned} & \left(z^2 + \left(\frac{p_0^2 + p_1^2 - \mu^2}{p_1 p_0} \right) z + 1 \right) h_0(z) \\ & = \frac{1}{p_0} p(z) u_1 + z m(z) \left(u_{-1} - \frac{1}{p_0} m(0) u_1 \right) \end{aligned} \quad (a3)$$

Recall that by Sarason's theorem $B_{opt} = \mu^2 q(T)^* h_0 / p(T)^* h_0$. In our case we then have

$$B_{opt}(z) = \mu^2 \frac{zh_0(z)}{(z p_0 + p_1) h_0(z) - p_1 u_1}$$

Set $\hat{p}(z) = z p_0 + p_1$, $\lambda(z) = (z^2 + (p_0^2 + p_1^2 - \mu^2)z / p_0 p_1 + 1)$, and $\hat{u}_{-1} = (u_{-1} - \frac{1}{p_0} m(0) u_1)$. Replacing $h_0(z)$ from (a3)

$$h_0(z) = \frac{p(z) u_1 / p_0 + z m(z) u_{-1}}{\lambda(z)} \quad (a3 a)$$

in $B_{opt}(z)$, and arranging terms we get

$$B_{opt}(z) = p(z) \frac{1 + m(z) z p_0 \hat{u}_{-1} / p(z) u_1}{1 + m(z) \hat{p}(z) p_0 \hat{u}_{-1} / \mu^2 u_1} \quad (a4)$$

On the other hand from equation (a3) we obtain that

$$\begin{aligned} & \frac{1}{p_0} p(\alpha_1) u_1 + \alpha_1 m(\alpha_1) \hat{u}_{-1} = 0 \\ & \frac{1}{p_0} p(\alpha_2) u_1 + \alpha_2 m(\alpha_2) \hat{u}_{-1} = 0 \end{aligned}$$

where α_1 and $\alpha_2 = \alpha_1^{-1}$ are the roots of $\lambda(z) = 0$ on the unit circle. So, we must have

$$p(\alpha_1) \alpha_2 m(\alpha_2) = p(\alpha_2) \alpha_1 m(\alpha_1) \quad (a5)$$

in order to have a nonzero vector h_0 satisfying $A_0 h_0 = 0$. Solving the equations, after tedious computations, one gets that $p_0 \hat{u}_{-1} / u_1 = \mu$. Putting this result in (a4), and solving for Q_{opt} from $B_{opt} = W - m Q_{opt}$, and then solving for the

optimum controller C_{opt} , via Youla parametrization, we end up with

$$C_{opt} = \left(\frac{W(z)(W(z)^{-1}) - 1}{\mu^2} \right) \frac{G(z)}{1 + G(z)m(z)} P_0^{-1}(z)$$

where $G(z) = \mu z / p(z)$. The computations are straightforward but too lengthy to present here.

When $m(s) = \frac{s-b}{s+h} e^{-\tau s}$ we find μ from the equation

$$h\omega_c + \tan^{-1} \tau_\omega \omega_c + \tan^{-1} \omega_c + 2 \tan^{-1} (\omega_c / b) = \pi$$

where

$$\omega_c = \frac{1}{\tau_\omega} \quad \text{and} \quad \tau_\omega = \sqrt{\frac{1}{\mu^2} - \frac{1}{\tau_\omega^2}}$$

This is obtained by writing the equation (a5) explicitly and transforming the data from the unit circle to right half plane using the transformation $z = \frac{\tau_\omega s + 1}{\tau_\omega s - 1}$. In the example considered in Section 6 when $\tau_\omega = \tau_\omega$, b and h are fixed ω_c and hence μ are found by using the above formula.

APPENDIX B

Proof of Condition (7)

Recall the expression

$$\hat{X}_1(s)$$

$$\frac{(1 - \tau_\omega \tau_\omega s - 1) - \tau_\omega \tau_\omega s \hat{m}(s) + \mu \frac{\tau_\omega s - 1}{\tau_\omega \tau_\omega s + 1} \omega_c \hat{m}(s) - \hat{m}(s)}{\mu \frac{\tau_\omega s + 1}{\tau_\omega \tau_\omega s + 1} - \tau_\omega \tau_\omega s + 1} \frac{\hat{m}(s)}{1 + \frac{1 - \tau_\omega \tau_\omega s - 1}{\mu \tau_\omega s + 1} \hat{m}(s)}$$

and definitions

$$\hat{R}_1(s) = 1 + \frac{1 - \tau_\omega \tau_\omega s - 1}{\mu \tau_\omega s + 1} \hat{m}(s),$$

$$D_1(s) = \hat{X}_1(s) \hat{R}_1(s)$$

Also from the equations (a3.a), (a4) and (a5) Appendix A it is easy to see that

$$\frac{1 - \tau_\omega \tau_\omega \omega_c}{\mu \tau_\omega \omega_c + 1} \hat{m}(j\omega_c) = -1 \quad (b1)$$

Moreover, $+j\omega_c$ and $-j\omega_c$ are the only points where $\hat{R}_1(s)$ vanishes in the closed right half plane. Therefore for the stability of \hat{X}_1 we need

$$D_1(j\omega_c) \neq 0 \quad (b2)$$

Using (b1) and re-arranging terms we get that (b2) is equivalent to having

$$(-1)(- \tau_\omega \tau_\omega \omega_c) + \left(\frac{1}{\hat{m}(j\omega_c)} \right) (\hat{m}_r(j\omega_c) - \hat{m}(j\omega_c)) = 0,$$

or

$$\hat{m}_r(j\omega_c) = \hat{m}(j\omega_c)(1 + \tau_\omega \tau_\omega \omega_c) \quad (b3)$$

It is also routine to check that

$$\frac{\partial}{\partial s} \left(\frac{\tau_\omega \tau_\omega s - 1}{\tau_\omega s + 1} \right) \hat{m}(s) \Big|_{s=j\omega_c} \neq 0 \quad (b4)$$

So, since $+j\omega_c$ and $-j\omega_c$ are the only points in the closed right half plane that makes $\hat{R}_1(s) = 0$, condition (b3) is also sufficient for the stability of \hat{X}_1 .

Now let's look at \hat{X}_2

$$\hat{X}_2 = \mu \frac{\tau_\omega s - 1}{\tau_\omega s + 1} \frac{1}{\tau_\omega \tau_\omega s + 1} \frac{\hat{m}(s)(- \tau_\omega \tau_\omega s) + (\hat{m}_r(s) - \hat{m}(s))}{1 + \frac{1 - \tau_\omega \tau_\omega s - 1}{\mu \tau_\omega s + 1} \hat{m}(s)}$$

From this expression, and the arguments used for \hat{X}_1 we see that the stability of \hat{X}_2 also is equivalent to (b3)

H^2 -optimal Control with an H^∞ -constraint: The State Feedback Case*

MARIO A. ROTE[†] and PRAMOD P. KHARGONEKAR[‡]

A state-space approach solves the problem of finding among all state feedback controllers that minimize an H^2 -performance measure one that also satisfies an H^∞ -norm bound.

Key Words—Linear optimal control, multiobjective optimization, robust control, state-feedback

Abstract—In this paper we consider a mixed H^2/H^∞ -optimal control problem. It is assumed that the plant as well as the feedback controller are finite-dimensional and linear time-invariant, and that the plant state is available for feedback. More specifically, among all the state-feedback controllers that minimize the H^2 -norm of a closed loop transfer matrix, we give necessary and sufficient conditions for the existence of a controller that also satisfies a prescribed H^∞ -norm bound on some other closed loop transfer matrix. When these conditions are met, the solution to the above problem is also a global solution to the constrained optimization problem of minimizing an H^2 -norm performance measure subject to an H^∞ -norm constraint. We also give state-space formulae for computing the solutions. Some easily checkable sufficient conditions for the existence of solutions are given. Finally we give an example in which all solutions to the constrained optimization problem are necessarily dynamic, i.e. there is no static gain solution even though plant state is available for feedback.

1. INTRODUCTION AND PROBLEM FORMULATION

THE CONTROL problem addressed in this paper concerns the finite-dimensional linear time-invariant feedback system depicted in Fig. 1. The signals w_1 and w_2 denote exogenous inputs while z_1 and z_2 denote controlled (i.e. regulated) signals. The signals u and y denote the control input and the measured output, respectively. The transfer matrices of the plant and the controller are denoted by G and K , respectively. It is also assumed that both G and K are real-rational and

proper transfer matrices. Finally, for given a real-rational and proper controller K , we let $T_1(K)$ and $T_2(K)$ denote the closed loop transfer matrices from w_1 to z_1 and w_2 to z_2 , respectively. When there is no possibility of confusion, the dependence of T_1 and T_2 on K will be omitted.

In this paper we assume that the state of the generalized plant G is available for feedback. To be more precise let a state-space description of G be given by:

$$\frac{dx}{dt} = Ax + B_1 w_1 + B_2 w_2 + B_3 u \quad (1.1a)$$

$$z_1 = C_1 x + D_1 u \quad (1.1b)$$

$$z_2 = C_2 x + D_2 u \quad (1.1c)$$

$$y = x \quad (1.1d)$$

where all the matrices in (1.1) are real matrices of compatible dimensions. Although no explicit frequency dependent weights were introduced, it is assumed that all weighting functions have been absorbed in the generalized plant G . Note that there are no feedthrough terms from the exogenous signals w to the controlled signals z . Although it is possible to include these terms, we have chosen not to include them in order to keep the presentation as simple as possible.

A given controller K is called *admissible* (for the plant G) if K is real-rational proper, and the minimal realization of K internally stabilizes the state-space realization (1.1) of G . Let $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the usual H^2 and H^∞ norms, respectively. The two problems considered in this paper are defined as follows:

Problem A: *Minimal H^2 -norm subject to an H^∞ -norm constraint.* For the plant G defined in (1.1), find an admissible controller K that

* Received 7 November 1989, revised 10 May 1990, received in final form 29 May 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor H. Kimura under the direction of Editor H. Kwakernaak.

[†] Center for Control Sciences and Dynamical Systems, University of Minnesota, Minneapolis, MN 55455, U.S.A. Now with the School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN 47907, U.S.A.

[‡] Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109-2122, U.S.A. Author to whom all correspondence should be addressed.

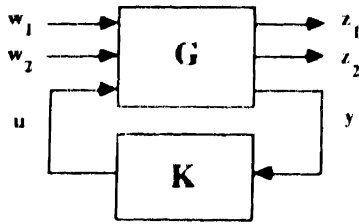


FIG. 1. The synthesis framework.

achieves

$$\inf \{ \|T_1(K)\|_2 : K \text{ admissible and } \|T_2(K)\|_2 < 1 \}.$$

Problem B: Simultaneous H^2/H^∞ optimal control. For the plant G defined in (1.1) find an admissible controller K that achieves

$$\inf \{ \|T_1(K)\|_2 : K \text{ admissible} \}$$

and such that $\|T_2(K)\|_2 < 1$.

Note that while Problem A represents a constrained optimization problem, Problem B is to find (if it exists) a solution to the unconstrained problem of minimizing an H^2 -performance measure that also satisfies an H^∞ -norm bound. The key point is that a solution to Problem B is also a solution to Problem A but the converse need not be true.

Recently Problem A has received a great deal of attention, mainly because it represents a problem of optimal nominal performance with robust stability [see, for example, Bernstein and Haddad (1989), Mustafa and Glover (1988), Doyle *et al.* (1989b)]. Indeed, if we consider that a stable (possibly nonlinear) perturbation Δ is connected from z_2 to w_2 (see Fig. 1), then the small gain theorem ensures stability of the perturbed system if the nominal system ($\Delta = 0$) is internally stable and $\|T_2\|_2 < 1$, provided that the induced operator norm of Δ is less than or equal to one. Among all the admissible controllers K that provide robust stability, Problem A is to find a controller that minimizes the variance of the output z_1 (with $\Delta = 0$) when w_1 is zero mean unit variance white noise.

Currently, no analytic solution to Problem A is known. Some attempts have been made to solve "modified" versions of this optimization problem. Mustafa and Glover (1988) and Glover and Mustafa (1989) have considered the special case in which $B_1 = B_2$, $C_1 = C_2$, $D_1 = D_2$, and hence $T_1 = T_2$ (see Fig. 1). For this case, they have solved the problem of maximizing an entropy functional subject to an H^∞ -norm constraint. This problem formulation is related to Problem A in that the (negative of the) entropy of a transfer matrix is an upper bound for its H^2 -norm. Bernstein and Haddad (1989)

have considered the case of one exogenous signal. In our setting, this means $B_1 = B_2$. They have also considered the minimization of an upper bound ("auxiliary cost", as defined by them) for $\|T_1\|_2$ subject to an H^∞ -norm constraint on T_2 . Using a Lagrange multiplier technique, and under the assumption that the order of the controller is specified, they have derived necessary conditions for optimality. See also Bernstein *et al.* (1989) for more recent work on this approach and Mustafa (1989) for an explicit connection between the entropy and the auxiliary cost for the special case $T_1 = T_2$. Finally, Doyle *et al.* (1989b) have considered a similar problem with one controlled output, i.e. $C_1 = C_2$ and $D_1 = D_2$. They have derived necessary conditions and sufficient conditions for this modified problem to have a solution. As shown in Doyle *et al.* (1989b), there may be a gap between these conditions. It is important to note that these papers address the more general and interesting situation of output feedback.

Problem B has not been considered before. Our objective in this paper is twofold. Firstly, we want to parametrize the set of all solutions for the (unconstrained) H^2 -optimal control problem $\inf \{ \|T_1(K)\|_2 : K \text{ admissible} \}$. Secondly, we want to find necessary and sufficient conditions for the existence of a solution to Problem B. Since a solution to Problem B is also a solution to Problem A, these conditions are sufficient for Problem A to have a solution. While it may seem that the solvability of Problem B is a very strong sufficient condition for the solvability of Problem A, it will be seen that if $\text{im}B_1$ and $\text{im}B_2$ are linearly independent, then Problem B and Problem A become equivalent.

It is important to note that the problems considered in this paper can also be approached with the aid of convex nonlinear programming; see, for example, Boyd *et al.* (1988). Since our results are analytical, they complement the numerical optimization approach taken by Boyd *et al.* (1988) and related papers.

A brief summary of our results and the organization of the paper is as follows. In Section 2 we give a parametrization of all H^2 -optimal state feedback controllers. This parametrization is obtained in terms of a free transfer matrix in RH^2 . Furthermore, any closed loop transfer matrix is *affine* in this free parameter. Using this parametrization along with the recent solution to the standard H^∞ problem given in Glover and Doyle (1988) and Doyle *et al.* (1989a), Section 3 gives necessary and sufficient conditions for the existence of solutions to Problem B (cf. Theorem 2). These

conditions involve two algebraic Riccati equations (AREs) and a coupling condition. The first ARE reflects the fact that there must exist an admissible (state-feedback) controller such that $\|T_2\|_\infty < 1$. The second ARE and the coupling condition arise due to the requirement of the H^2 optimization. Since a solution to Problem B is also a solution to Problem A, these conditions are sufficient for Problem A to have a solution (cf. Corollary 1). When these conditions are satisfied, a "dynamic" state-feedback controller that solves Problem B (and Problem A) is given. Finally, in Section 4, we consider the special case in which $\text{im}B_1$ and $\text{im}B_2$ are linearly independent. In this case we show that Problem A has a solution if and only if there exists an admissible controller such that $\|T_2\|_\infty < 1$. Thus, under the mild condition of linear independence of $\text{im}B_1$ and $\text{im}B_2$, one gets a complete solution to Problem A. In this section we also show (by example) that there are situations in which any solution to either Problem A or Problem B must necessarily be dynamic. This is in significant contrast to the fact that in either H^2 or H^∞ optimal control problems, when states are available for feedback, the controller can be chosen to be a memoryless gain [see, for example Kalman (1960) and Khargonekar *et al.* (1988)]. This example appears to indicate that the mixed H^2/H^∞ problems are likely to be much more complicated than standard H^2 and H^∞ problems. Finally in Section 5 the conclusions of this work are given.

The notation is fairly standard. The identity matrix is denoted by I . For a constant matrix M , let $\text{im}M$ and $\text{ker}M$ denote its range and null space, respectively. The spectral radius of M is denoted by $\rho(M)$. The transpose of M is denoted by M' . Let M^+ denote the Moore-Penrose inverse of M . If $M = 0$ we shall define $M^+ := 0$. The orthogonal complement of a subspace $S \subset \mathbb{R}^p$ is denoted by S^\perp . Packed matrix notation is used to represent state-space realizations, i.e.

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} := C(sI - A)^{-1}B + D.$$

For a transfer matrix G , we define \bar{G} as $\bar{G}(s) := G'(-s)$ for all complex s . The spaces H^2 and H^{2*} denote the Hardy spaces of matrix-valued functions that are square integrable on the imaginary axis with analytic extensions into the right and left half plane, respectively. The Hardy space H^∞ consists of matrix-valued functions that are bounded on the imaginary axis with analytic extension into the right half plane. The norms in these spaces are defined in the

usual way:

$$\|G\|_2 := \sqrt{\frac{1}{2\pi} \int_{-\infty}^{\infty} \text{trace}(G - \bar{G})(j\omega) d\omega}$$

$$\|G\|_\infty := \sup_{\omega \in \mathbb{R}} \sigma\{G(j\omega)\}, \quad (\sigma := \text{max singular value}).$$

When R is used as a prefix, it denotes real rational. Given real matrices A , $R = R'$, and $Q = Q'$, will say that the algebraic Riccati equation (ARE) $A'X + XA + XRX + Q = 0$ admits the (unique) stabilizing solution X , if X is real and symmetric, X satisfies the ARE, and $A + RX$ has eigenvalues in the open left half plane. A similar definition is used for the antistabilizing solution with the obvious modifications.

In the rest of the paper, the following standard assumptions on the plant G are made:

- (i) $\{A, B_k\}$ is stabilizable. (1.2a)
- (ii) D_1 and D_2 both have full column rank. (1.2b)
- (iii) For each $\omega \in \mathbb{R}$ and for $k = 1, 2$

$$\begin{bmatrix} A - j\omega I & B_k \\ C_k & D_k \end{bmatrix} \quad (1.2c)$$

has full column rank

Note that assumption (1.2a) ensures that the set of admissible controllers is non-empty while assumptions (1.2b, c) are standard and guarantee that the LQR problem corresponding to the quadratic cost $\|z_k\|_2$ has an admissible solution. Without loss of generality we further assume that

$$(iv) \quad D_2^+ [C_2^+ D_2] = [0 \quad I]. \quad (1.2d)$$

In fact, as is well known, a preliminary feedback transformation will enforce this last equation.

2. PARAMETRIZATION OF ALL H^∞ -OPTIMAL CONTROLLERS

In this section we parametrize the set of all admissible unconstrained H^2 -optimal state-feedback controllers for the interconnection of Fig. 1. The development is carried out using a frequency domain approach. More specifically, the YJBK parametrization of all stabilizing compensators is used to solve this problem (see for example Vidyasagar, 1984). The final formula for the solution to this problem is given in state-space (cf. Theorem 1).

Consider the block diagram of Fig. 1, where the plant G is as in (1.1). It is a classical fact that [under assumptions (1.2a-c)] there exists an admissible controller that minimizes $\|T_1\|_2$. One

such admissible controller is $K_0 = F$, where the constant real matrix F is computed according to

$$F := -(D_1' D_1)^{-1} (D_1' C_1 + B_1' X_1), \quad (2.1a)$$

and the constant matrix X_1 is the unique stabilizing solution of the (LQR) ARE

$$A'X + XA - (D_1' C_1 + B_1' X)(D_1' D_1)^{-1} \times (D_1' C_1 + B_1' X) + C_1' C_1 = 0. \quad (2.1b)$$

Perhaps, it is less well known that (2.1) is not the only admissible controller that minimizes $\|T_1\|_2$. Our first result (Theorem 1) gives a complete parametrization of all solutions to this optimization problem. With reference to the realization of the plant G given in (1.1) and with F given by (2.1), define

$$\Pi_1 := I - B_1 B_1', \quad (2.2)$$

$$A_F := A + B_1 F, \quad (2.3)$$

$$C_{kF} := C_k + D_k F; \quad k = 1, 2. \quad (2.4)$$

The matrix Π_1 is the orthogonal projection onto $(\text{im } B_1)^\perp$. Note that A_F is a stability matrix. Define the set of transfer matrices:

$$S := \{Q \in RH^+ : Q = W \Pi_1 (sI - A_F), W \in RH^2\}. \quad (2.5)$$

Theorem 1. Consider the feedback system of Fig. 1, with the plant G given by (1.1). Let S be defined by (2.5). Let K denote an admissible controller and T_1 the corresponding closed loop transfer matrix from w_1 to z_1 . Then, K minimizes $\|T_1\|_2$ if and only if K equals the transfer matrix from y to u in

$$J := \begin{bmatrix} A_F & 0 & B_1 \\ 0 & F & I \\ -I & I & 0 \end{bmatrix} \quad (2.6)$$

for some $Q \in S$.

Note that if $\text{im } B_1 = \mathbf{R}^n$ (n := state dimension) then $\Pi_1 = 0$ and (2.7) reduces to the single state-feedback controller $K_0 = F$. On the other hand, if $\text{im } B_1$ is a proper subspace of \mathbf{R}^n then (2.6) generates a family of controllers parametrized by W . This extra freedom can be used to satisfy some additional constraints.

The next lemma will be useful for establishing Theorem 1. It provides state-space formulae for the YJBK parametrization of all admissible controllers. The formulae given below are more appropriate for our setting (state-feedback) than

the well known formulae in terms of an "observer-based" stabilizing compensator.

Lemma 1. Consider the feedback system of Fig. 1, where G is given by (1.1). Then, a given controller K is admissible if and only if there exists $Q \in RH^+$ such that K equals the transfer matrix from y to u in (2.6).

Proof. First, note that a given controller K is admissible for G if and only if K is admissible for $G_{yu} := (sI - A)^{-1} B_1$. With F given by (2.1), define the RH^+ matrices

$$N := \begin{bmatrix} A_F & B_1 \\ I & 0 \end{bmatrix},$$

$$M := I + FN, \quad \tilde{N} := N, \quad \tilde{M} := I + NF,$$

$$X := F, \quad Y := I, \quad \tilde{X} := F, \quad \tilde{Y} := I.$$

Then, it is straightforward to verify that

$$G_{yu} = NM^{-1} = \tilde{M}^{-1} \tilde{N}$$

$$\begin{bmatrix} \tilde{Y} & -\tilde{X} \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & X \\ N & Y \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

These equations provide a (doubly) coprime factorization over RH^+ for G_{yu} . It now follows (see, for example, Vidyasagar, 1984) that K is admissible for G_{yu} if and only if there exists $Q \in RH^+$ such that

$$K = (X + MQ)(Y + NQ)^{-1} \\ = F + Q(I + NQ)^{-1},$$

which is of the form (2.6). Q.E.D.

Proof of Theorem 1. From Lemma 1 and after some standard algebraic manipulations, it follows that the set of all admissible closed loop transfer matrices from w_1 to z_1 (i.e. those that are generated by admissible controllers) can be parametrized by the formula

$$T_1 := S_1 + U_1 Q V_1, \quad Q \in RH^+, \quad (2.7)$$

$$S_1 := \begin{bmatrix} A_F & B_1 \\ C_{1F} & 0 \end{bmatrix}, \quad U_1 := \begin{bmatrix} A_F & B_1 \\ C_{1F} & D_1 \end{bmatrix},$$

$$V_1 := \begin{bmatrix} A_F & B_1 \\ I & 0 \end{bmatrix},$$

where A_F and C_{1F} are defined in (2.3) and (2.4), respectively.

It is now standard to show (using some simple algebra and (2.1)) that $R_1 := D_1' D_1 = U_1' U_1$ and that $U_1 S_1 \in H^{2 \times 1}$. Using these properties, we conclude from (2.9) that if $Q \in RH^+$ then

$$\|T_1\|_2^2 = \|S_1\|_2^2 + \|U_1 Q V_1\|_2^2 \\ = \|S_1\|_2^2 + \|\sqrt{R_1} Q V_1\|_2^2.$$

It is now clear that $Q \in RH^\infty$ minimizes $\|T_1\|_2$ if and only if Q satisfies

$$\sqrt{R_1} Q V_1 = 0 \Leftrightarrow Q(sI - A_F)^{-1} B_1 = 0. \quad (2.8)$$

To complete the proof we must show that $Q \in RH^\infty$ satisfies (2.8) if and only if $Q \in S$. Clearly, if $Q \in S$ then $Q \in RH^\infty$ and satisfies (2.8). The converse is as follows. Let $Q \in RH^\infty$ be given and suppose that it satisfies (2.8). Define $W := Q(sI - A_F)^{-1}$. Note that $W \in RH^2$, for A_F is a stability matrix. Moreover, from (2.2) and (2.8), it follows that

$$W \Pi_1 = Q(sI - A_F)^{-1} \Rightarrow W \Pi_1 (sI - A_F) = Q.$$

Therefore, $Q \in S$. Q.E.D.

We conclude this section with a state-space representation for the controller of Theorem 1 that will be useful for establishing the main result of this paper (Theorem 2). Let $W \in RH^2$ be given by

$$W = \frac{A_\kappa \mid B_\kappa}{C_\kappa \mid 0}$$

Then it is easy to show that $Q = W \Pi_1 (sI - A_F)$ is given by

$$Q = \frac{A_\kappa \mid A_\kappa B_\kappa \Pi_1 - B_\kappa \Pi_1 A_F}{C_\kappa \mid C_\kappa B_\kappa \Pi_1}$$

Substituting this realization of Q in (2.6), and after deleting unobservable modes, the controller K of Theorem 1 is given by

$$A_K = A_\kappa - B_\kappa \Pi_1 B_1 C_\kappa, \quad (2.9a)$$

$$K = \frac{A_\kappa \mid A_\kappa B_\kappa \Pi_1 - B_\kappa \Pi_1 A_F}{C_\kappa \mid F + C_\kappa B_\kappa \Pi_1} \quad (2.9b)$$

3. THE SIMULTANEOUS H^2/H^∞ PROBLEM

In this section we solve the simultaneous H^2/H^∞ optimization problem (Problem B) defined in Section 1. The development is carried out using Theorem 1 along with the recent solution to the standard H^∞ -optimization problem given in Glover and Doyle (1988) and Doyle *et al.* (1989a). For the sake of completeness, a slightly modified statement (suitable for our purposes) of the main result of Glover and Doyle (1988) has been included in a separate appendix.

Our first result in this section gives necessary and sufficient conditions for the existence of solutions to Problem B. Consider the plant G given by (1.1) and let Π_1 denote the projection matrix defined in (2.2). Define also

$$V_2 := \Pi_1 B_2. \quad (3.1)$$

Note that $V_2 = 0$ if and only if $\text{im} B_2 \subset \text{im} B_1$. The

following algebraic Riccati equations (for X and Y) will play an important role in stating our conditions for the existence of solutions to Problem B:

$$A'X + XA + X(B_2 B_2' - B_1 B_1')X + C_2' C_2 = 0 \quad (3.2)$$

$$YA_F' + A_F Y + Y(C_2' C_2)Y + B_2(I - V_2' V_2)B_2' = 0 \quad (3.3)$$

where F , A_F , and C_2 , are defined in (2.1), (2.3) and (2.4), respectively.

Theorem 2. Consider the feedback system of Fig. 1, with the plant G given by (1.1). Then there exists an admissible controller K that solves Problem B if and only if the following conditions hold:

(i) The ARE (3.2) admits the stabilizing solution X_2 , and $X_2 \neq 0$. (3.5a)

(ii) The ARE (3.3) admits the stabilizing solution Y_2 . (3.5b)

(iii) $\rho(Y_2 X_2) < 1$. (3.5c)

Moreover, when these conditions are met, one solution to Problem B is given by

$$\frac{A_0 \mid A_0 \Sigma - \Sigma A_F}{H - F \mid F(I - \Sigma) + H \Sigma}, \quad (3.6a)$$

where

$$A_0 := A + (I - \Sigma)B_1 H + \Sigma B_1 F + (I - \Sigma)B_2 B_2' X_2, \quad (3.6b)$$

$$\Sigma := Z_2 B_2 V_2' \Pi_1, \quad H := -B_1' X_2, \quad Z_2 := (I - Y_2 X_2)^{-1}, \quad (3.6c)$$

and F , A_F are defined in (2.1) and (2.3), respectively.

It should be noted that, under assumption (1.2d), condition (3.5a) is equivalent to the existence of an admissible controller K such that $\|T_2\|_\infty < 1$ (Doyle *et al.*, 1989a). The other two conditions, namely (3.5b, c), reflect the fact that one of these admissible controllers must also minimize $\|T_1\|_2$. As it will become clear from the proof of Theorem 2, when $\text{im} B_2 \subset \text{im} B_1$, Problem B has a solution if and only if condition (3.5b) holds. In this case, the constant gain F defined in (2.1) is a solution to Problem B.

Since a solution to Problem B is also a solution to Problem A, Theorem 2 may be used to produce a sufficient condition under which Problem A can be solved. It is obvious that Problem A makes sense only if condition (3.5a) holds. In other words, if (3.5a) is not satisfied then, there is no state-feedback controller that internally stabilizes the plant G and yields

$\|T_2\|_\infty < 1$. An immediate consequence of Theorem 2 is the following.

Corollary 1. Consider the feedback system of Fig. 1, with the plant G given by (1.1). Suppose that conditions (3.5) hold. Then the controller given by (3.6) is an admissible controller that solves Problem A. Furthermore,

$$\inf \{ \|T_1\|_2 : K \text{ admissible and } \|T_2\|_\infty < 1 \} \\ = \inf \{ \|T_1\|_2 : K \text{ admissible} \}.$$

We conclude this section with a number of intermediate results that will lead to a proof of Theorem 2. In the light of Theorem 1, solving Problem B reduces to that of finding a controller of the form (2.7) such that $\|T_2\|_\infty < 1$. Let F , A_F , and C_{2F} be defined by (2.1), (2.3) and (2.4), respectively. An easy calculation shows that when the controller given in Theorem 1 is connected to the plant G , the closed loop transfer matrix from w_2 to z_2 equals

$$T_2 := S_2 + U_2 W V_2, \quad W \in RH^2 \text{ (free)}, \quad (3.8a)$$

where

$$S_2 := \frac{A_F}{C_{2F}} \frac{B_2}{0}, \quad U_2 := \frac{A_F}{C_{2F}} \frac{B_1}{D_2}, \quad V_2 := \Pi_1 B_2. \quad (3.8b)$$

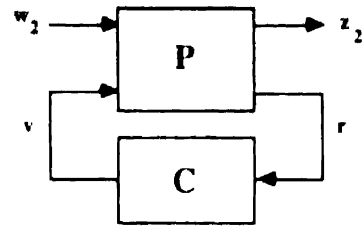
From (3.8b) we observe that if $\text{im} B_2 \subset \text{im} B_1$ then $V_2 = 0$, and $T_2 = S_2$ for all $W \in RH^2$. Therefore, in this particular case, we can not exploit the transfer matrix W to reduce $\|T_2\|_\infty$. In other words, for all $W \in RH^2$ it follows that $T_2 = T_2(F)$, where F is the constant gain given in (2.1). The following lemma gives necessary and sufficient conditions for the existence of $W \in RH^2$ such that $\|T_2\|_\infty < 1$ and it will become useful for establishing Theorem 2.

Lemma 2. Consider the transfer matrix T_2 given in (3.8). Assume that $V_2 \neq 0$. Then, there exists $W \in RH^2$ such that $\|T_2\|_\infty < 1$ if and only if conditions (3.5) hold. In this case, a transfer matrix $W \in RH^2$ such that $\|T_2\|_\infty < 1$ is given by

$$W = \frac{A_H + (I - \Sigma)B_2 B_2' X_2}{H - F} \left| \frac{Z_2 B_2 V_2'}{0} \right|, \quad (3.10)$$

where $A_H := A + B_1 H$, and H , Σ and Z_2 are defined in (3.6c).

Proof. First we factor the non-zero constant matrix V_2 as $V_2 = M_0 M_1$, where M_0 is a full column rank matrix and M_1 satisfies $M_1 M_1' = I$. (Note that this factorization always exists.) Now it is easy to see that the closed loop transfer matrix T_2 given in (3.8) equals the transfer matrix from w_2 to z_2 in the following diagram:



$$P := \left[\begin{array}{c|cc} A_F & B_2 & B_1 \\ \hline C_{2F} & 0 & D_2 \\ \hline 0 & M_1 & 0 \end{array} \right] \quad C := W M_0. \quad (3.11)$$

Note that the full column rank property of M_0 guarantees that the existence of $W \in RH^2$ such that $\|T_2\|_\infty < 1$ is equivalent to the existence of $C \in RH^2$ such that $\|T_2\|_\infty < 1$.

Next, we show that conditions (3.5) are necessary and sufficient for the existence of such a transfer matrix C . First note that since the open loop transfer matrix P_{∞} (from v to r) in (3.11) is identically zero, and since A_F is a stability matrix, it is obvious that a given controller C is admissible for P if and only if $C \in RH^2$. We claim that the auxiliary plant P defined in (3.11) satisfies all the assumptions of Theorem A.7 (see the Appendix). This claim will be verified later.

Now, applying the result of Theorem A.7 to the auxiliary plant P and after some algebra, one concludes that there exists $C \in RH^2$ such that $\|T_2\|_\infty < 1$ if and only if the following conditions are met:

- The ARE (3.2) admits the stabilizing solution X_2 , and $X_2 \geq 0$. [Here we have used assumption (1.2d).]
- The ARE (3.3) admits the stabilizing solution Y_2 , and $Y_2 \geq 0$. [Here we have used the identity $M_1' M_1 = V_2' V_2$.]
- $\rho(Y_2 X_2) < 1$.

The equivalence between the condition (b) above and (3.5b) is obtained by observing that the stability of A_F implies that any symmetric solution to the ARE (3.3) is positive semidefinite. We must also show that when the above conditions are met there is a choice of C not only in RH^2 but also in RH^1 . This immediately follows from the construction of the controller given in Theorem A.7. In fact, from (A.8a), we observe that C can be chosen to be strictly proper.

Finally, assuming that conditions (3.5) hold, the formula for W given in (3.10) follows from (A.8) (to obtain a formula for C), and solving the linear equation indicated in (3.11) for the transfer matrix W . In this step we have used the fact that there always exist a choice of M_1 and M_0 such that $V_2' = M_1' M_0^-$ (where M_0^- denotes a left inverse of M_0).

To complete the proof, we must verify that the auxiliary plant P in (3.11) satisfies all the assumptions of Theorem A.7. Clearly, (A.1) follows from the stability property of A_F . Assumption (A.2) follows from (1.2d) and the identity $M_i M_i' = I$. Finally, using (1.2c) and the identity

$$\begin{bmatrix} A_F - j\omega I & B_1 \\ C_{2F} & D_2 \end{bmatrix} = \begin{bmatrix} A - j\omega I & B_1 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ F & I \end{bmatrix},$$

we conclude that (A.3) is satisfied. The fact that assumption (A.4) holds is a consequence of the nonsingularity of A_F and the full row rank property of M_i . Q.E.D.

Proof of Theorem 2. We consider two cases.

Case 1. Suppose that $\text{im} B_2 \not\subset \text{im} B_1$. In this case $V_2 \neq 0$ (cf. (3.1)). Combining the results of Theorem 1 and Lemma 2 it is clear that among the admissible controllers that minimize $\|T_i\|_2$ there exists one such that $\|T_2\|_2 < 1$ if and only if conditions (3.5) hold. Suppose that these conditions hold and let W be given by (3.10). Define

$$A_w := A_H + (I - \Sigma)B_2 B_2' X_2,$$

$$B_w := Z_2 B_2 V_2', \quad C_w := H - F.$$

Using these equations and the fact that $\Sigma := B_w \Pi_1$ [cf. (3.6c)], it follows from (2.11) that a solution to Problem B is given by (3.6).

Case 2. Suppose that $\text{im} B_2 \subset \text{im} B_1$. In this case V_2 and V_2' are both zero [cf. (3.1)]. Therefore, from (3.8a), we conclude that Problem B has a solution if and only if

$$\|S_2\|_2 < 1, \quad (3.12)$$

where the transfer matrix S_2 is defined in (3.8b). We claim that (3.12) and (3.5b) are equivalent. In fact, since A_F is a stability matrix, from Lemma 4 in Doyle *et al.* (1989a) it follows that (3.12) is satisfied if and only if the ARE

$$Y A_F' + A_F Y + Y(C_{2F}' C_{2F})Y + B_2 B_2' = 0, \quad (3.13)$$

admits the stabilizing solution. The claim is finally established by observing that $V_2' V_2 = 0$ implies that the AREs (3.3) and (3.13) are the same. Next, we show that conditions (3.5a) and (3.5c) are necessary for Problem B to have a solution.

Suppose that Problem B has a solution; then (3.12) holds. From (3.8a) it follows that the admissible controller F defined in (2.1) yields the closed loop transfer matrix $T_2(F) = S_2$. Hence, from (3.12) and item F1.4 in Doyle *et al.* (1989a), we conclude that condition (3.5a) must hold.

The necessity of condition (3.5c) will be

proved under the technical assumption that the pair (C_2, A) is observable. (Even if this is not the case, the result is still true, and can be obtained by factoring the unobservable subspace out.) Suppose that Problem B has a solution, then conditions (3.5a, b) are satisfied. It is easy to show [using the observability of the pair (C_2, A)] that the stabilizing solution to the ARE (3.2) satisfies $X_2 > 0$. Define $Y_* := X_2^{-1}$. From (3.2) it follows that Y_* satisfies

$$A Y_* + Y_* A' + Y_* C_2' C_2 Y_* + B_2 B_2' - B_1 B_1' = 0,$$

$$(A + Y_* C_2' C_2)'$$

$$= -Y_*^{-1}(A + (B_2 B_2' - B_1 B_1')X_2)Y_*, \quad (3.14)$$

Hence, Y_* is the (unique) *anti-stabilizing* solution to the ARE (3.14). We now show that if Y denotes any real symmetric solution to the ARE (3.3) then

$$Y_* \preceq Y. \quad (3.15)$$

Indeed, using a "completion of squares" argument, it follows that any solution Y of (3.3) satisfies the quadratic matrix inequality

$$\begin{aligned} A Y + Y A' + Y C_2' C_2 Y + B_2 B_2' - B_1 B_1' \\ = -(B_1' + F Y)'(B_1' + F Y) \preceq 0. \end{aligned}$$

Hence, from Ran and Vreugdenhil (1988), it follows that inequality (3.15) must hold.

Note that the pair (C_{2F}, A_F) is observable. This follows from the observability of (C_2, A) and assumption (1.2d). Since (3.5b) holds we conclude from Willems (1971) that the anti-stabilizing solution to the ARE (3.3), say Y_* , exists and

$$Y_* \preceq Y_2, \quad (3.16)$$

where Y_2 denotes the stabilizing solution to the ARE (3.3). Combining (3.15) and (3.16) we obtain that $Y_* \preceq Y \preceq Y_2$, which implies that $X_2^{-1} \preceq Y_2$. Therefore, condition (3.5c) is satisfied.

To complete the proof we must show that when conditions (3.5) hold, the controller K in (3.6) solves Problem B. This is immediate since $V_2' = 0$ implies that $\Sigma = 0$ [cf. (3.6c)]. Thus (3.6a) reduces to

$$K = \begin{bmatrix} A + (B_2 B_2' - B_1 B_1')X_2 & 0 \\ H - F & F \end{bmatrix} = F.$$

Since F is an admissible controller, the result follows from (3.12) and the fact that $T_2(F) = S_2$.

Q.E.D.

4. SPECIAL CASES

In this section we will focus on Problem A. Corollary 1 tells us that if conditions (3.5) hold,

then there is a solution to the unconstrained problem $\inf \{ \|T_1(K)\|_2 : K \text{ admissible} \}$ that also solves Problem A. Therefore, one might be tempted to conclude that these conditions are too restrictive. We claim that this is not the case. In this section we show that if $\text{im}B_1 \cap \text{im}B_2 = 0$, then conditions (3.5b, c) hold. In fact, under this geometric condition, a much stronger result is true. As before, let $\Pi_1 := I - B_1 B_1^+$ and $V_2 := \Pi_1 B_2$.

(4.1)

Lemma 3. Consider the feedback system of Fig. 1, with the plant G given by (1.1). Assume that $\text{im}B_1 \cap \text{im}B_2 = 0$. Let F_1 and F_2 denote two (arbitrary) constant state-feedback matrices for the plant G . Then the dynamic state-feedback controller

$$K := \begin{bmatrix} A_1 & A_1 \Delta - \Delta A_{F_1} \\ F_2 - F_1 & F_1(I - \Delta) + F_2 \Delta \end{bmatrix} \quad (4.2a)$$

where

$$A_1 := A + (I - \Delta)B_1 F_2 + \Delta B_1 F_1, \quad (4.2b)$$

$$A_{F_1} := A + B_1 F_1, \quad \Delta := B_2 V_2^+ \Pi_1, \quad (4.2c)$$

achieves the following closed loop transfer matrices: $T_1(K) = T_1(F_1)$ and $T_2(K) = T_2(F_2)$. Moreover, K is admissible if and only if both F_1 and F_2 are admissible.

Proof. Note that $\Delta B_1 = 0$, since Π_1 is the orthogonal projection onto $(\text{im}B_1)^\perp$. Next we show that $\Delta B_2 = B_2$. It is easy to see that $\text{im}B_1 \cap \text{im}B_2 = 0$ implies that $\ker V_2 = \ker B_2$. Since $V_2^+ V_2$ and $B_2^+ B_2$ are the orthogonal projections onto $(\ker V_2)^\perp$ and $(\ker B_2)^\perp$ respectively, and since orthogonal projections are unique, we conclude that

$$V_2^+ V_2 = B_2^+ B_2.$$

Therefore,

$$\Delta B_2 = B_2 V_2^+ V_2 = B_2 B_2^+ B_2 = B_2.$$

(Actually, Δ is a projection onto $\text{im}B_2$.)

Let the controller K be given by (4.2). Let x and ξ denote the states of G and K respectively. Consider the interconnection of Fig. 1 and define new coordinates x_n and ξ_n according to $x_n := (I - \Delta)x - \xi$ and $\xi_n := \Delta x + \xi$. It is easy to verify that the transformation from (x, ξ) to (x_n, ξ_n) is invertible. Using the fact that $\Delta B_1 = 0$ and $\Delta B_2 = B_2$, a trivial computation shows that the closed loop equation for the block diagram of Fig. 1 is given by

$$\frac{dx_n}{dt} = (A + B_1 F_1)x_n + B_1 w_1$$

$$\frac{d\xi_n}{dt} = (A + B_2 F_2)\xi_n + B_2 w_2$$

$$z_1 = (C_1 + D_1 F_1)x_n + (C_1 + D_1 F_2)\xi_n$$

$$z_2 = (C_2 + D_2 F_1)x_n + (C_2 + D_2 F_2)\xi_n$$

which completes the proof.

Q.E.D.

Lemma 3 tells us that whenever the two subspaces $\text{im}B_1$ and $\text{im}B_2$ are independent, there exists a dynamic state-feedback controller that simultaneously achieves closed loop transfer matrices T_1 and T_2 , provided there exist constant state feedback matrices F_1 and F_2 such that $T_1 = T_1(F_1)$ and $T_2 = T_2(F_2)$. The connection between Lemma 3 and Problem A is given by the following lemma.

Lemma 4. Consider the feedback system of Fig. 1, with the plant G given by (1.1). Suppose that $\text{im}B_1 \cap \text{im}B_2 = 0$. Then Problem A is solvable if and only if there exists an admissible controller K such that $\|T_2(K)\|_\infty < 1$. In this case, a solution to Problem A is given by (4.2) with $F_1 := F$ and $F_2 := H$, where F and H are defined in (2.1) and (3.6c), respectively.

Proof. The necessity immediately follows from the definition of Problem A. The sufficiency part is as follows. First, note that the existence of an admissible controller K such that $\|T_2(K)\|_\infty < 1$ implies that condition (3.5a) holds. In this case, the constant matrix H in (3.6c) is an admissible controller such that $\|T_2(H)\|_\infty < 1$. (See Doyle *et al.*, 1989a). Note also that the constant matrix F in (2.1) is an admissible controller that minimizes $\|T_1\|_2$. Choosing $F_1 := F$ and $F_2 := H$, the result follows from Lemma 3. Q.E.D.

It is also interesting to establish a connection between Lemma 4 and Corollary 1. Suppose that $\text{im}B_1 \cap \text{im}B_2 = 0$. We will show that conditions (3.5b, c) are automatically satisfied. Recall from the proof of Lemma 3 that under this geometric condition, $V_2^+ V_2 = B_2^+ B_2$. Hence, the ARE (3.3) reduces to

$$Y A_F^* + A_F Y + Y (C_F^* C_F) Y = 0, \quad (4.3)$$

which has the unique stabilizing solution $Y_2 = 0$. (Recall that A_F is a stability matrix.) Thus, conditions (3.5b, c) hold and from Corollary 1 we conclude that Problem A has a solution if and only if condition (3.5a) holds. In this case, a trivial computation shows that the controllers of Corollary 1 and Lemma 4 are the same.

We conclude this section with an example. Our objective in this example is to show that although Problem A need not have a unique solution, there are situations in which any solution must necessarily be "dynamic". With reference to Fig. 1, let G be given by [note that

G satisfies Assumptions (1.2))

$$G := \begin{array}{c} \begin{array}{c} z_1 \\ z_2 \\ y \end{array} \left\{ \begin{array}{c|cc} \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix} & \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ \hline \begin{bmatrix} 0 & 0 \\ \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \\ \hline \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{array} \right. \end{array} \quad (4.4)$$

First, note that the ARE (3.2) admits the stabilizing solution

$$X_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

which is positive definite. Thus, the set of admissible controllers K such that $\|T_2\|_\infty < 1$ is non-empty. Since $\text{im} B_1 \cap \text{im} B_2 = 0$, Lemma 4 may be used to solve Problem A. From (2.1) it follows that $F = [0 \ 0]$. This is obvious since $z_1 = u$ and the plant G is stable. Next, we compute the gain $H := -B_1^T X_2$. In this case we obtain $H = -[0.5 \ 0.5]$. Finally, from (4.2) it follows that a solution to Problem A is given by

$$K = \begin{array}{c|cc} \begin{array}{cc} -1 & 1 \\ 0 & -1 \end{array} & \begin{array}{cc} 1 & 1 \end{array} \\ \hline \begin{array}{cc} -0.5 & -0.5 \end{array} & \begin{array}{cc} 0.5 & -0.5 \end{array} \end{array}$$

which obviously has McMillan degree equal to two. We now show that in this example any other solution to Problem A must be dynamic. First, note that

$\inf \{ \|T_1(K)\|_\infty : K \text{ admissible and}$

$$\|T_2(K)\|_\infty < 1 \} = 0.$$

Using some simple algebra it is easily seen that the unique state feedback "gain" that achieves this optimal performance is $F = [0 \ 0]$. From (4.4) it follows that $\|T_2(F)\|_\infty = \sqrt{2}$. Thus we conclude that for this particular example any solution to Problem A must be necessarily dynamic.

5 CONCLUSIONS

For the state-feedback case we have completely solved a mixed H^2/H^∞ control problem (Problem B). Necessary and sufficient conditions for the existence of solutions to Problem B were given in terms of solutions to certain AREs. A closed form expression for a solution was also provided. A solution to Problem B (when it exists) also solves the constrained optimal control problem of minimizing an H^2 performance measure subject to an H^∞ constraint

(Problem A). This problem is well motivated since it models a problem of optimal nominal performance with robust stability. Previous authors have only considered the minimization of an upper bound for the H^2 design objective. In this sense, the results of this work constitute the first results on this problem.

From Lemma 3 it follows that if the two subspaces $\text{im} B_1$ and $\text{im} B_2$ are independent, then one can always find a dynamic state-feedback controller that simultaneously achieves given closed loop transfer matrices T_1 and T_2 provided they can be separately achieved using static state-feedback controllers. The simplest case for which the condition of independence of $\text{im} B_1$ and $\text{im} B_2$ is not satisfied is when $B_1 = B_2$. Recall from the proof of Theorem 2 that in this case Problem B has a solution if and only if $\|T_2(F)\|_\infty < 1$, where F denotes the LQR gain defined in (2.1). Therefore Problem B does not help much in solving Problem A. In this sense, further research on Problem A for the case $B_1 = B_2$ should be most useful.

The example in Section 4 illustrates that, even though the plant state is available for feedback, Problems A and B need not have a static solution. This is in significant contrast to the classical results in the LQR theory (Kalman, 1960) and the recent results in H^∞ control theory (Khargonekar *et al.*, 1988) which show that these optimal control problems always have a static state-feedback solution. This may have some implications in the output-feedback case. For instance, it might turn out that in the output-feedback case the dimension of optimal controllers in mixed H^2/H^∞ problems exceeds the plant dimension.

Acknowledgements—This research was supported in part by NSF under grant no. ECS-9096109, AFOSR under contract no. AFOSR-90-0053, and ARO under grant no. DAA103-90-G-0008. The first author was also supported by the Graduate School Doctoral Dissertation Fellowship, University of Minnesota.

REFERENCES

- Bernstein, D. S. and W. M. Haddad (1989) LQG control with an H^∞ performance bound: A Riccati equation approach. *IEEE Trans. Aut. Control*, **AC-34**, 293.
- Bernstein, D. S., W. M. Haddad and C. N. Nett (1989) Minimal complexity control law synthesis, part 2: problem solution via H^2/H^∞ optimal static output feedback. *Proc. 1989 Amer. Control Conf.*, Pittsburgh, PA, 2506.
- Boyd, S. P., V. Balakrishnan, C. H. Barratt, N. M. Khraishi, X. Li, D. G. Meyer and S. A. Norman (1988) A new CAD method and associated architectures for linear controllers. *IEEE Trans. Aut. Control*, **AC-33**, 268.
- Doyle, J. C., K. Glover, P. P. Khargonekar and B. A. Francis (1989a) State-space solutions to standard H^2 and H^∞ control problems. *IEEE Trans. Aut. Control*, **AC-34**, 831.
- Doyle, J. C., K. Zhou and B. Bodenheimer (1989b) Optimal control with mixed H^2 and H^∞ performance

objectives. *Proc. 1989 Amer. Control Conf.*, Pittsburgh, PA, 2065.

Glover, K. and J. C. Doyle (1988). State-space formulae for all stabilizing controllers that satisfy an H^∞ norm bound and relations to risk sensitivity. *Syst. Control Lett.*, **11**, 167.

Glover, K. and D. Mustafa (1989). Derivation of the maximum entropy H^∞ -controller and a state-space formula for its entropy. *Int. J. Control*, **50**, 899.

Kalman, R. E. (1960). Contributions to the theory of optimal control. *Bol. Soc. Mat. Mexico*, **5**, 102.

Khargonekar, P. P., I. R. Petersen and M. A. Rotea (1988). H^∞ -optimal control with state-feedback. *IEEE Trans. Aut. Control*, **AC-33**, 786-788.

Mustafa, D. (1989). Relations between maximum-entropy/ H^∞ control and combined H^∞ /LQG control. *Syst. Control Lett.*, **12**, 193.

Mustafa, D. and K. Glover (1988). Controllers which satisfy an H^∞ -norm bound and maximize an entropy integral. *Proc. 1988 CDC*, Austin, TX, 959.

Ran, A. C. M. and R. Vreugdenhil (1988). Existence and comparison theorems for algebraic Riccati equations for continuous and discrete time systems. *Linear Algebra Applic.*, **99**, 63.

Vidyasagar, M. (1984). *Control Systems Synthesis. A Factorization Approach*, MIT Press, Cambridge, MA.

Willems, J. C. (1971). Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Trans. Aut. Control*, **AC-16**, 621.

APPENDIX

State-space formulae for the standard H^∞ problem

The result given in this appendix provides a solution to the standard H^∞ -optimal control problem, and is a slight modification of Theorem 1 in Glover and Doyle (1988). Consider the feedback system shown in Fig. 2, where both the plant P and the controller C are real-rational and proper. Let $T(C)$ denote the closed loop transfer matrix from w to z . Assume that P has the following realization:

$$\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_1 \\ \hline C_2 & D_2 & 0 \end{array}$$

along with the assumptions

- (i) (A, B_2) stabilizable and (C_2, A) detectable. (A.1)
- (ii) $D_1^T D_1 \neq I$ and $D_2^T D_2 = I$. (A.2)
- (iii) For each $\omega \in \mathbb{R}$

$$\begin{array}{c|c} A - j\omega I & B_2 \\ \hline D_1 & \end{array} \quad (\text{A.3})$$

has full column rank

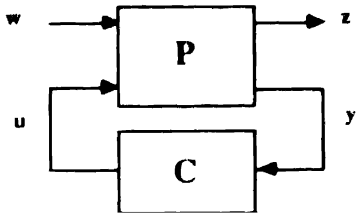


FIG. 2. Feedback system.

(iv) For each $\omega \in \mathbb{R}$

$$\begin{array}{c|c} A - j\omega I & B_1 \\ \hline C_2 & D_2 \end{array} \quad (\text{A.4})$$

has full row rank.
Now, define the matrices

$$A_1 := A - B_2 D_1^T C_1, \quad A_2 := A - B_1 D_2^T C_2.$$

In the result given below we will make use of the following algebraic Riccati equations for X and Y , respectively.

$$A_1^T X + X A_1 + X (B_1 B_1^T - B_2 B_2^T) X + C_1^T (I - D_1 D_1^T) C_1 = 0 \quad (\text{A.5})$$

$$A_2 Y + Y A_2 + Y (C_1^T C_1 - C_2^T C_2) Y + B_1 (I - D_2^T D_2) B_1^T = 0. \quad (\text{A.6})$$

The next result is a slightly modified version of Theorem 1 in Glover and Doyle (1988).

Theorem A.7 Consider the feedback system of Fig. 2. Then there exists an admissible controller C such that $\|T(C)\|_\infty \leq 1$ if and only if the following conditions are satisfied:

- (i) The ARE (A.5) admits the stabilizing solution X_+ , and $X_+ \geq 0$.
- (ii) The ARE (A.6) admits the stabilizing solution Y_+ , and $Y_+ \leq 0$.
- (iii) $\rho(Y_+ X_+) < 1$.

Moreover, when these conditions hold one such a controller is

$$C := \begin{bmatrix} A_{F_+} + (B_1 + Z_+ L_+ D_2) B_1^T X_+ + Z_+ L_+ C_2 & -Z_+ L_+ \\ F_+ & 0 \end{bmatrix}, \quad (\text{A.7})$$

where

$$\begin{aligned} L_+ &= -(Y_+ C_2^T + B_1 D_2^T), \quad F_+ = -(B_2^T X_+ + D_1^T C_1), \\ A_{F_+} &= A + B_2 F_+, \quad Z_+ := (I - Y_+ X_+)^{-1}. \end{aligned} \quad (\text{A.8b})$$

ℓ^1 -optimal Control of Multivariable Systems with Output Norm Constraints*

J. S. McDONALD† and J. B. PEARSON†‡

The ℓ^1 -optimal control problem is considered for general rational plants, possibly subject to ℓ^∞ -norm constraints on some outputs, and a procedure given for the construction of optimal or near-optimal rational compensators.

Key Words—Control system synthesis; linear optimal control; multivariable control systems; linear programming.

Abstract—In this paper, we consider the ℓ^1 -optimal control problem for general rational plants. It is shown that for plants with no poles or zeros on the unit circle an optimal compensator exists and that the resulting closed loop transfer function is polynomial whenever there are at least as many controls as regulated outputs and at least as many measurements as exogenous inputs. Exactly or approximately optimal rational compensators can be obtained by solving a sequence of finite linear programs for the coefficients of a polynomial closed loop transfer function. No assumptions on plant poles or zeros are required to obtain at least approximately optimal compensators. It is shown that constrained problems in which a set of outputs is regulated subject to ℓ^∞ -norm constraints on another set of outputs can be solved using a slight modification of the same algorithm.

Notation

Let m and n be positive integers. Then we define:

$\ell_{m \times n}^1$ The real normed linear space of all $m \times n$ matrices \hat{H} each of whose entries is a right-sided, absolutely summable real sequence $\hat{H}_{ij} = (\hat{H}_{ij}(k))_{k=0}^\infty$. The norm is defined:

$$\|\hat{H}\|_1 := \max_{i \in \{1, \dots, m\}} \sum_{j \in \{1, \dots, n\}} \sum_{k=0}^\infty |\hat{H}_{ij}(k)|$$

$\ell_{m \times n}^\infty$ The real normed linear space of all $m \times n$ matrices \hat{H} each of whose entries is a right-sided, magnitude bounded real sequence $\hat{H}_{ij} = (\hat{H}_{ij}(k))_{k=0}^\infty$. The norm is defined:

$$\|\hat{H}\|_\infty := \sum_{i=1}^m \max_{j \in \{1, \dots, n\}} \sup_k |\hat{H}_{ij}(k)|$$

$\ell_{m \times n}^0$ The subspace of $\ell_{m \times n}^1$ consisting of all elements

each of whose entries converges to zero, that is:

$$\ell_{m \times n}^0 := \left\{ \hat{H} \in \ell_{m \times n}^1 : \lim_{k \rightarrow \infty} \hat{H}_{ij}(k) = 0 \right. \\ \left. \forall i \in \{1, \dots, m\}, j \in \{1, \dots, n\} \right\}$$

Let m and n be as above and let z be a complex variable. Then we define

D, \bar{D} The open and closed, respectively, unit disk in the complex plane
 $\mathcal{F}(\cdot)$ The \mathcal{F} -transform. Given a matrix $\hat{H} = (\hat{H}_{ij}(k))_{k=0}^\infty$ of right-sided real sequences

$$\mathcal{F}(\hat{H}) := \sum_{k=0}^\infty \hat{H}(k)z^k$$

$A_{m \times n}$ The real normed linear space of all $m \times n$ matrices H such that $H = \mathcal{F}(\hat{H})$ for some sequence $\hat{H} \in \ell_{m \times n}^1$. The norm is defined $\|H\|_A := \|\hat{H}\|_1$.
 $\mathcal{H}A_{m \times n}$ The subspace of $A_{m \times n}$ consisting of all elements each of whose entries is a real-rational function of z . (Entries are precisely those with all poles outside \bar{D} .)

We will often drop the m and n in the above notations when the dimension is either unimportant or clear from the context. Because the \mathcal{F} -transform is an invertible mapping on all the above sequence spaces, we can associate sequences and their \mathcal{F} -transforms as pairs. For such a pair, we write a hatted variable to denote the sequence and an unhatted variable to denote the corresponding \mathcal{F} -transform.

Now let X be a real normed linear space, let $S \subset X$ be a subspace of X , and let x^* be a bounded linear functional on X . Then we define

BX The closed unit ball of X , $BX := \{x \in X : \|x\| \leq 1\}$.
 $\langle x, x^* \rangle$ The value of the bounded linear functional x^* at the point $x \in X$.
 X^* The space of all bounded linear functionals on X , also called the dual space of X . X^* is a complete real normed linear space with the norm defined:

$$\|x^*\| := \sup_{x \in BX} |\langle x, x^* \rangle|$$

S^\perp The (right) annihilator of $S \subset X$; $S^\perp := \{x^* \in X^* : \langle x, x^* \rangle = 0 \forall x \in S\}$. Thus S^\perp is a subspace of X^* .
 ${}^{\perp}S$ The (left) annihilator of $S \subset X$; ${}^{\perp}S := \{x \in X : \langle x, x^* \rangle = 0 \forall x^* \in S\}$. Thus ${}^{\perp}S$ is a subspace of X .

* Received 11 September 1989; revised 26 April 1990; received in final form 31 May 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor H. Kimura under the direction of Editor H. Kwakernaak.

† Department of Electrical and Computer Engineering, Rice University, Houston, TX 77251-1892, U.S.A.

‡ Author to whom all correspondence should be addressed.

1. INTRODUCTION

In this paper, we consider the problem of minimizing the maximum magnitude of a set of regulated outputs of a linear discrete time system excited by bounded-amplitude disturbance signals, using a linear shift invariant discrete-time compensator. This is equivalent to minimization of the A -norm of a closed loop transfer matrix of the form, given a transfer matrix $H \in \mathcal{H}A$,

$$\Phi = H - K$$

where K takes values in some feasible set \mathcal{F} . The required compensator is constructed from K . The feasible set depends on two given transfer matrices U and V in $\mathcal{H}A$. Taking $\mathcal{F} = S := \{K \in \mathcal{H}A : \exists Q \in \mathcal{H}A \text{ satisfying } K = UQV\}$ corresponds to minimizing over all stabilizing compensators with rational transfer matrices, while taking $\mathcal{F} = S_A := \{K \in A : \exists Q \in A \text{ satisfying } K = UQV\}$ corresponds to allowing stabilizing compensators with transfer matrices in the quotient field of A . From a practical standpoint, the former version of the problem is most useful since it considers only finite dimensional compensators, while the latter allows the compensator to be possibly infinite dimensional. However, because of the properties of A as the dual of a normed linear space, the latter problem has been the standard one studied.

In Dahleh and Pearson (1987), the problem with $\mathcal{F} = S_A$ was considered under several assumptions on U and V : that U has full row rank, V has full column rank, neither U nor V have transmission zeros on the unit circle, all the zeros of U and V in D are simple, and U and V have no common zeros in D . The rank assumption on U and V corresponds to requiring that the open loop system have at least as many independent control inputs as outputs to be regulated and at least as many independent measured outputs as disturbance inputs, while the zeros of U and V arise from the poles and zeros of the open loop system. It was shown that a $K_0 \in S_A$ exists which minimizes $\|\Phi\|_A$ and that, in fact, any such K_0 must be in S so that the problem has a solution when $\mathcal{F} = S$. Moreover for the minimizing K_0 , $\Phi_0 = H - K_0$ is a polynomial which can be constructed from the solution of a sufficiently large finite linear program formulated in a dual space. This linear program is equivalent to the problem with the feasible set restricted to include only K s such that $\Phi = H - K$ is a polynomial of fixed degree at least equal to the degree of Φ_0 .

In Dahleh and Pearson (1988), a particular problem in which U has dimensions 2×1 and V has dimensions 1×2 was considered (that is, the above rank assumptions are not satisfied).

Similar assumptions were made on the zeros of U and V , entrywise. It was shown once again that a minimizing K exists in S_A , but without the above rank assumptions it is unclear whether a minimizer exists in S . It was shown, however, that if there exists a $K \in S_A$ such that $\Phi = H - K$ is polynomial, then a sequence of increasingly large finite linear programs can be formulated with the property that the minimum norms form a non-increasing sequence which converges to the infimal norm. Thus this gives a method of finding approximate minimizers which are arbitrarily good by solving a sufficiently large linear program. Each linear program corresponds to restricting the feasible set to K s such that $\Phi = H - K$ is a polynomial of a fixed finite degree. Again, these linear programs are formulated in a dual space and the corresponding Φ s constructed from their solutions.

In this paper, we have two main aims. First, we wish to give a method for computing implementable compensators with optimal or at least close to optimal performance in the most general possible setting. In this spirit, we will take $\mathcal{F} = S$ in the formulation of our standard problem and drop as many of the above assumptions on U and V as possible. Second, we address the problem of minimizing the maximum magnitude of a set of regulated outputs subject to constraints on the maximum magnitude of other outputs. This corresponds to a "disk" type constraint on the norm of Φ .

In Section 2 we discuss briefly the formulation of our standard problem and identify four cases determined by the rank (i.e. row or column) of the matrices U and V . These cases are treated separately throughout the paper and lead to significant differences in the properties of the problem.

In Section 3, we use essentially the same approach as Dahleh and Pearson (1987, 1988) to address existence, that is, we consider the problem first with $\mathcal{F} = S_A$ and then infer existence of a minimizer in S if possible. We will retain only the assumption that U and V have no unit circle zeros and show that a minimizer exists in S_A regardless of the ranks of U and V . In the case in which U has full row rank and V has full column rank, a minimizer exists in S and corresponds to a polynomial Φ . This is the expected generalization of the results of Dahleh and Pearson. The characterization of the feasible set S_A in this general case is the main new result in this section and provides the key not only to proving existence but to computing minimizers. Also, our proofs of existence of minimizers are somewhat more direct than those previously given.

In Section 4, we drop all assumptions on U and V and consider the computation of exact or approximate minimizers in S . First we note that the characterization found in Section 3 of the feasible set S_A extends in an obvious way to characterize S even when unit circle zeros are present in U and/or V . We show that when U has full row rank and V has full column rank, arbitrarily good approximate minimizers can always be found by solving finite linear programs, which again correspond to Φ being a polynomial of a fixed degree. We formulate the linear program in terms of the coefficients of the polynomial Φ . Of course, when U and V have no unit circle zeros, a sufficiently large such program will yield a minimizing K . When U and V do not satisfy the rank assumptions, there may be no $K \in S$ which gives Φ polynomial, so that the above procedure cannot be used. We give a simple characterization of all problems which have this difficulty.

Finally, in Section 5, we define a class of constrained problems as described above. We do not consider the existence of minimizers for such problems, but we show that, provided the constrained problem is feasible, arbitrarily good approximate minimizers can be obtained as easily as for the unconstrained case by a slight modification of the same algorithm.

2. PROBLEM FORMULATION

In the standard problem which we will consider, we are given a discrete time system G which is causal, finite dimensional linear and shift invariant (FDLSI) and hence is described by a real-rational transfer matrix $G(z)$. The system G has two (vector) inputs and two (vector) outputs: w is a vector of n_w exogenous disturbance inputs, z is a vector of n_z outputs which are to be regulated, y is a vector of n_y outputs which are measurable, and u is a vector of n_u control inputs. The control input u is assumed to be the output of a causal FDLSI compensator C whose input is the measured output y . If we partition the transfer matrix $G(z)$ conformally with these inputs and outputs, we obtain the following equations to describe the closed loop system:

$$\begin{pmatrix} z \\ y \end{pmatrix} = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \begin{pmatrix} w \\ u \end{pmatrix} \quad (2.1)$$

$$u = Cy.$$

It is well known that if the system G is admissible (Cheng and Pearson, 1981), the set of all closed loop transfer matrices Φ from w to z which may be achieved by choice of C , and which correspond to an internally stable

closed-loop system, may be described by the following parametrization (see e.g. Francis, 1987):

$$\Phi = H - UQV \quad (2.2)$$

where:

$$H := G_{11} + G_{12}M\hat{Y}G_{21}$$

$$U := G_{12}M$$

$$V := \hat{M}G_{21}$$

$$Q \in \mathcal{RA}$$

and $G_{22} = NM^{-1} = \hat{M}^{-1}\hat{N}$ are arbitrary right and left stable coprime factorizations of G_{22} (i.e. coprime over \mathcal{RA}). Also, the following Bezout identity is satisfied:

$$\begin{pmatrix} \hat{X} & -\hat{Y} \\ -\hat{N} & \hat{M} \end{pmatrix} \begin{pmatrix} M & Y \\ N & X \end{pmatrix} = I.$$

The transfer matrices H , U , and V are all in \mathcal{RA} and the transfer matrix Q is the free design parameter. The $\Phi \in \mathcal{RA}$ corresponding to a particular Q may be achieved by choosing the compensator transfer matrix C as follows:

$$C = (Y - MQ)(X - NQ)^{-1} \quad (2.3a)$$

$$= (\hat{X} - Q\hat{N})^{-1}(Y - Q\hat{M}). \quad (2.3b)$$

In this setting the problem of minimization of $\|\hat{z}\|$, when $\hat{w} \in BC'$ is equivalent to the following minimum distance problem in \mathcal{RA} :

$$(OPT): \quad \inf_{K \in S} \|H - K\| = \mu_{opt}$$

where $S := \{K \in \mathcal{RA} : \exists Q \in \mathcal{RA} \text{ satisfying } K = UQV\}$. We will sometimes say, given a problem (OPT), that (OPT) is the standard problem defined by H , U and V given above. Also, given a system G' and input and output weighting transfer matrices W_i , $W_o \in \mathcal{RA}$, the problem of minimizing $\|\hat{W}_i * \hat{z}\|$, when the inputs are in a weighted ball $B_{W_o} := \{\hat{w} \in C' : \exists \hat{y} \in BC' \text{ satisfying } \hat{w} = W_o \hat{y}\}$ may be put in the above form by taking $H = W_i H'$, $U = W_i U'$, and $V = V' W_o$.

The only assumption which will be considered to hold generally throughout the paper is the following:

Assumption 1. The transfer matrices U and V have full normal rank, that is, full rank for almost all z .

Special cases

Given the above assumption, the dimensions of the inputs and outputs of the standard system of (2.1) determine what kind of rank (i.e. full row rank or full column rank or both) that U and V have, and we can classify problems (OPT) by rank as follows. In the case $n_u \geq n_z$, or at least as

many control inputs as regulated outputs, U will be a "fat" matrix with full row rank. If we have $n_y \geq n_u$, or at least as many measured outputs as exogenous inputs, V will be a "skinny" matrix with full column rank. Intuitively, this combination of dimensions is good from the point of view of designing compensators for the system, and the study of (OPT) turns out to be simpler in this case. For this reason we will call this the *good rank* case. If $n_u < n_x$, U will be a skinny matrix with full column rank and we will say this U has *bad rank*. Similarly, $n_y < n_u$ will result in a fat V with full row rank, and we will say this V has bad rank. Problems in which U and/or V have bad rank will be called bad rank problems.

The various combinations of good and bad rank in U and V thus define four cases of (OPT). Note that for the H^* problem, it is also these combinations of ranks in U and V which define the usual four cases of that problem (the one-block problem, the two cases of two-block problems, and the four-block problem). We will consider in detail just the two extreme cases in which either both U and V have good rank (the one-block problem) or both have bad rank (the four-block problem). The intermediate cases (two-block problems) have properties essentially like the latter, and we will make comments in the sequel to indicate how they can be treated.

4. EXISTENCE OF A MINIMIZER

In this section, we consider the question: When does there exist $K_0 \in S$ such that $\mu_{\text{opt}} = \|H - K_0\|_A$? Our approach is to make use of the following two theorems from Luenberger (1969), which ensure existence of solutions to minimum distance problems set in the duals of normed linear spaces and identify a useful property called alignment of such solutions when they exist.

Definition 1. Given a real normed linear space X and its dual X^* , we say that an element $x^* \in X^*$ and an element $x \in X$ are *aligned* if $\langle x, x^* \rangle = \|x\| \|x^*\|$.

Theorem 1. Let x be an element in a real normed linear space X and let d denote its distance from the subspace M . Then:

$$d = \inf_{m \in M} \|x - m\| = \max_{x^* \in B M^\circ} \langle x, x^* \rangle$$

where the maximum on the right is achieved for some $x_0^* \in M^\circ$ with $\|x_0^*\| = 1$. If the infimum on the left is achieved for some $m_0 \in M$, then $x - m_0$ is aligned with x_0^* .

Theorem 2. Let M be a subspace in a real normed linear space X . Let $x^* \in X^*$ and let d denote its distance from M° . Then:

$$d = \min_{m^* \in M^\circ} \|x^* - m^*\| = \sup_{x \in B M} \langle x, x^* \rangle$$

where the minimum on the left is achieved for some $m_0^* \in M^\circ$. If the supremum on the right is achieved for some $x_0 \in M$, then $x^* - m_0^*$ is aligned with x_0 .

The following corollary to Theorem 2 follows easily using the fact that if a subspace M in a dual space is *weak**-closed then $M = [^\perp M]^\circ$ (Rudin, 1973, Thm 4.7). In fact, it is equivalent to Theorem 2 (except for the alignment condition) using the fact that for every subspace M of a normed linear space X , M° is *weak**-closed in X^* (Rudin, 1973, p. 91).

Corollary 1. Let X be a normed linear space. Let $x^* \in X^*$ and let M^* be a subspace of X^* . If M^* is *weak**-closed, then there exists an element $m_0^* \in M^*$ such that:

$$\|x^*\| = \inf_{m^* \in M^*} \|x^* - m^*\|$$

Since the space $\mathcal{H}A$ is not complete and hence cannot be a dual space, the question of existence of a minimizer for (OPT) cannot be resolved directly using these results. Instead we consider the related minimum distance problem in $\ell_{n_x \times n_u}^1$:

$$(OPT_1): \quad \inf_{K \in S_1} \|\hat{H} - \hat{K}\|_1 =: \mu_1$$

where $S_1 := \{\hat{K} \in \ell_{n_x \times n_u}^1 : \exists Q \in A \text{ satisfying } K = UQV\}$. We will use the facts (Luenberger, 1969) that $(\ell_{m \times n}^0)^* = \ell_{m \times n}^1$ when we define linear functional evaluation as follows, given $\hat{G} \in \ell_{m \times n}^0$ and $\hat{H} \in \ell_{m \times n}^1$:

$$\langle \hat{G}, \hat{H} \rangle := \sum_{i=1}^m \sum_{j=1}^n \sum_{k=0}^{\infty} \hat{G}_{ij}(k) \hat{H}_{ij}(k)$$

and that $(\ell_{m \times n}^1)^* = \ell_{m \times n}^0$ with a similar definition of functional evaluation.

(OPT₁) is the problem which was considered in Dahleh and Pearson (1987, 1988). It is also clearly equivalent to (OPT) with the feasible set enlarged from just S to all of S_A , since $A_{n_x \times n_u}$ and $\ell_{n_x \times n_u}^1$ are isometrically isomorphic under the \mathcal{Z} -transform. Our approach will be to use Theorem 2 to establish first the existence of a minimizer for (OPT₁) (which corresponds to a point in S_A). The alignment condition of Theorem 1 will then allow us to conclude in certain cases that the \mathcal{Z} -transform of this

minimizer lies, in fact, in S and hence is a minimizer for (OPT).

We will see that the results of Dahleh and Pearson (1987, 1988) concerning existence extend to the general case in the expected way; in the good rank case, we will be able to establish existence of a minimizer for (OPT) assuming no zeros of U and V are on the unit circle while in the bad rank case, we will need a similar assumption and will only be able to establish existence of a minimizer for (OPT₁). As has been shown by counterexample in Vidyasagar (1987) we cannot hope to establish existence in general without precluding unit circle zeros.

The good rank case

In this case, U has full row rank $= n_r$ and V has full column rank $= n_v$. Before proceeding, we establish some notation. For simplicity, we replace the dimensions n_r and n_v with m and n , respectively. We will need to consider Smith-McMillan form decompositions (MacFarlane and Karcanas, 1976) of U and V given by:

$$\begin{aligned} U &= L_U M_U R_U \\ V &= L_V M_V R_V \end{aligned} \quad (3.1)$$

where L_U , R_U , L_V , and R_V are (polynomial) unimodular matrices and M_U , M_V are rational matrices which have the familiar diagonal forms:

$$\begin{aligned} M_U &= \begin{pmatrix} \frac{\epsilon_1(z)}{\psi_1(z)} & & 0 & \cdots & 0 \\ & \ddots & \vdots & \ddots & \vdots \\ & & \frac{\epsilon_m(z)}{\psi_m(z)} & & 0 \\ & & & & 0 \end{pmatrix}; \\ M_V &= \begin{pmatrix} \frac{\epsilon'_1(z)}{\psi'_1(z)} & & & & \\ & \ddots & & & \\ & & \frac{\epsilon'_n(z)}{\psi'_n(z)} & & \\ 0 & \cdots & & 0 & \\ \vdots & \vdots & & \vdots & \\ 0 & \cdots & & 0 & \end{pmatrix} \end{aligned} \quad (3.2)$$

Let \mathcal{Z}_{UV} denote the set of all $z \in \hat{D}$ which are zeros of either U or V . Then for each $z_0 \in \mathcal{Z}_{UV}$ we can define a non-decreasing sequence of non-negative integers $\Sigma_U(z_0)$ corresponding to the multiplicities with which the term $(z - z_0)$ appears on the diagonal of M_U . That is:

$$\Sigma_U(z_0) := (\sigma_{U_i}(z_0))_{i=1}^m$$

means:

$$\frac{\epsilon_i(z)}{\psi_i(z)} = (z - z_0)^{\sigma_{U_i}(z_0)} g_i(z) \quad i = 1, \dots, m$$

where the $g_i(z)$ have no poles or zeros at $z = z_0$. We can define similarly a set of sequences $\Sigma_V(z_0)$

for each $z_0 \in \mathcal{Z}_{UV}$ which correspond to the multiplicities of the z_0 s on the diagonal of M_V . A sequence $\Sigma_U(z_0)$ is sometimes referred to as the sequence of *structural indices* of z_0 in U .

We can also define m polynomial row vectors of dimension m and n polynomial column vectors of dimension n as follows:

$$\begin{aligned} \alpha_i(z) &= (L_U^{-1})_i(z) \quad i = 1, \dots, m \\ \beta_j(z) &= (R_V^{-1})^j(z) \quad j = 1, \dots, n \end{aligned}$$

where subscript i indicates the i th row and superscript j indicates the j th column. We can now state the assumption we will require and define a set of conditions which will be shown to characterize the feasible set S_1 of (OPT₁).

Assumption 2. Neither U nor V have any transmission zeros on the unit circle, that is, $\mathcal{Z}_{UV} \subset D$.

Definition 2. Given U and V as above and $K \in A_{m \times n}$, we say K interpolates U (from the left) and V (from the right) if the following condition is satisfied:

Given any zero $z_0 \in \mathcal{Z}_{UV}$ of U and/or V with structural indices $\Sigma_U(z_0)$ and $\Sigma_V(z_0)$ in U and V , respectively, we have for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$:

$$(i) \quad (\alpha_i K)^{(k)}(z_0) = 0, \quad k = 0, \dots, \sigma_{U_i} - 1,$$

$$(ii) \quad (K \beta_j)^{(k)}(z_0) = 0, \quad k = 0, \dots, \sigma_{V_j} - 1,$$

$$(iii)(a) \quad \sum_{l=0}^k \sum_{r=0}^{\sigma_{U_i}-1} \binom{k}{l} \binom{k-l}{r} \times [\alpha_i^{(l)} K^{(k-l-r)} \beta_j^{(r)}](z_0) = 0,$$

$$k = \sigma_{U_i}, \dots, \sigma_{U_i} + \sigma_{V_j} - 1$$

or: (b)

$$\sum_{l=0}^k \sum_{r=0}^{\sigma_{U_i}-1} \binom{k}{l} \binom{k-l}{r} \times [\alpha_i^{(l)} K^{(k-l-r)} \beta_j^{(r)}](z_0) = 0,$$

$$k = \sigma_{U_i}, \dots, \sigma_{V_j} + \sigma_{U_i} - 1$$

where the argument of $\sigma_{U_i}(\cdot)$ and $\sigma_{V_j}(\cdot)$ is understood to be z_0 and superscript (k) indicates the k th derivative with respect to z .

Note that this condition simplifies greatly in the case of a zero z_0 which is not common to U and V ; if it is a zero only of U , for example, we have $\Sigma_V(z_0) = (0)_{j=1}^n$ and parts (ii) and (iii) are trivially satisfied for all i and j . The following theorem shows that this condition characterizes

S_i in terms of its image in A under the \mathcal{T} -transform.

Theorem 3. Given Assumption 2, U and V as above, and $K \in A$, there exists $Q \in A$ satisfying $K = UQV$ if and only if K interpolates U and V .

We defer the proof of the theorem until we have established the following two lemmas.

Lemma 1. Given Assumption 2, U and V as above, and $K \in A_{m \times n}$ there exists $Q \in A_{m \times n}$ satisfying $K = UQV$ if and only if for all $z_0 \in \mathcal{T}_{UV}$, $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ we have:

$$(\alpha_i K \beta_j)^{(k)}(z_0) = 0, \\ k = 0, \dots, \sigma_{U_i}(z_0) + \sigma_{V_j}(z_0) - 1.$$

Proof. We can clearly factor M_U and M_V of (3.2) into forms:

$$M_U = M_{U_1} M_{U_2}, \quad M_V = M_{V_1} M_{V_2}$$

where $M_{U_1} = \text{diag}[(\lambda_{U_i})_{i=1}^m]$ and $M_{V_1} = \text{diag}[(\lambda_{V_j})_{j=1}^n]$ are diagonal matrices containing exactly the zeros in D of M_U and M_V , respectively, and M_{U_2} , M_{V_2} are in A with right and left inverses, respectively, in A . Hence:

$$\begin{aligned} \exists Q \in A \quad \text{satisfying} \quad K &= UQV \\ \Updownarrow \\ \exists \tilde{Q} \in A \quad \text{satisfying} \quad K &= L_U M_{U_2} \tilde{Q} M_{V_2} R_V \\ \Updownarrow \\ \tilde{Q} &:= M_{U_2}^{-1} L_U^{-1} K R_V^{-1} M_{V_2}^{-1} \in A. \end{aligned}$$

But this last holds if and only if an arbitrary entry of \tilde{Q} is in A , i.e. if and only if:

$$\tilde{Q}_{ij} = \frac{\alpha_i K \beta_j}{\lambda_{U_i} \lambda_{V_j}} \in A \tag{*}$$

for i, j arbitrary. To show that this is equivalent to the condition in the lemma, first suppose (*) holds. Then $\alpha_i K \beta_j = \lambda_{U_i} \lambda_{V_j} \tilde{Q}_{ij}$ where K and \tilde{Q} are both analytic in D . Thus the conditions in the lemma hold (Churchill and Brown, 1984, p. 152). Conversely, if the conditions in the lemma hold, it can be shown that we can write $\alpha_i K \beta_j = \lambda_{U_i} \lambda_{V_j} \tilde{K}_{ij}$ where $\tilde{K}_{ij} \in A$. Hence $\tilde{Q}_{ij} = \tilde{K}_{ij} \in A$ and (*) holds. \square

Lemma 2. Given non-negative integers σ_U , σ_V , and $k \leq \sigma_U + \sigma_V - 1$, $K \in A$, polynomial row and column vectors α and β , respectively, and

$$\begin{aligned} z_0 \in \bar{D}: \\ (\alpha K \beta)^{(k)}(z_0) &= \sum_{l=0}^{k-\sigma_U} \binom{k}{l} (\alpha K)^{(l)}(z_0) \beta^{(k-l)}(z_0) \\ &\quad + \sum_{l=0}^{\sigma_U-1} \binom{k}{l} \alpha^{(k-l)}(z_0) (K \beta)^{(l)}(z_0) \\ &\quad + \sum_{l=0}^{k-\sigma_U-\sigma_V+1} \sum_{r=0}^{\sigma_V-1} \binom{k}{l} \binom{k-l}{r} \alpha^{(l)}(z_0) \\ &\quad \times K^{(k-l-r)}(z_0) \beta^{(r)}(z_0) \\ &= \sum_{l=0}^{\sigma_U-1} \binom{k}{l} (\alpha K)^{(l)}(z_0) \beta^{(k-l)}(z_0) \\ &\quad + \sum_{l=0}^k \binom{k}{l} \alpha^{(k-l)}(z_0) (K \beta)^{(l)}(z_0) \\ &\quad + \sum_{l=0}^{k-\sigma_U} \sum_{r=0}^{\sigma_V-1} \binom{k}{l} \binom{k-l}{r} \alpha^{(l)}(z_0) \\ &\quad \times K^{(k-l-r)}(z_0) \beta^{(r)}(z_0) \end{aligned}$$

Proof. This follows straightforwardly, if somewhat tediously, by manipulating the expansion:

$$(\alpha K \beta)^{(k)} = \sum_{l=0}^k \sum_{r=0}^l \binom{k}{l} \binom{l}{r} \alpha^{(r)} K^{(l-r)} \beta^{(k-l)}$$

and using the fact that α , β , and K are all in A . \square

Proof of Theorem 3. Using Lemma 1, we see that to prove the theorem we can equivalently establish for all $z_0 \in \mathcal{T}_{UV}$, $i \in \{1, \dots, m\}$, and $j \in \{1, \dots, n\}$:

(i), (ii), and (iii) of Definition 2 hold

$$\begin{aligned} (\alpha_i K \beta_j)^{(k)}(z_0) &= 0, \\ k &= 0, \dots, \sigma_{U_i}(z_0) + \sigma_{V_j}(z_0) - 1. \end{aligned} \quad \Updownarrow$$

(\Rightarrow): This follows immediately by applying Lemma 2 in the appropriate form.

(\Leftarrow): Let $z_0 \in \mathcal{T}_{UV}$ be arbitrary. If we establish that (i) holds for arbitrary i and (ii) holds for arbitrary j , (iii) follows by applying Lemma 2. Considering (i) first, let i be arbitrary and argue inductively on k . For $k = 0$ we have:

$$\begin{aligned} 0 &= (\alpha_i K \beta_j)(z_0) = (\alpha_i K)(z_0) \beta_j(z_0), \\ &\quad \forall j \in \{1, \dots, n\} \\ &\Rightarrow (\alpha_i K)(z_0) R_V^{-1}(z_0) = 0 \\ &\Rightarrow (\alpha_i K)(z_0) = 0 \end{aligned}$$

since R_V^{-1} is unimodular. Now suppose $(\alpha_i K)^{(l)}(z_0) = 0$ for $0 \leq l \leq k-1$. Then we have

for all j :

$$\begin{aligned} 0 &= (\alpha, K\beta_j)^{(k)}(z_0) \\ &= \sum_{l=0}^k \binom{k}{l} (\alpha, K)^{(l)}(z_0) \beta_j^{(k-l)}(z_0) \\ &= (\alpha, K)^{(k)}(z_0) \beta_j(z_0) \\ &\quad + \sum_{l=0}^{k-1} \binom{k}{l} (\alpha, K)^{(l)}(z_0) \beta_j^{(k-l)}(z_0) \\ &= (\alpha, K)^{(k)}(z_0) \beta_j(z_0) \end{aligned}$$

since the summation on the right is zero. Thus we have:

$$(\alpha, K)^{(k)}(z_0) R_1^{-1}(z_0) = 0 \Rightarrow (\alpha, K)^{(k)}(z_0) = 0.$$

Hence, (i) holds. The argument to establish (ii) is similar. \square

We should note that while Lemma 1 provides a notationally simpler set of conditions equivalent to that of Theorem 3, the latter provides a smaller *number* of conditions on K . It will be clear from the next section that this will lead to fewer constraint equations in a linear program which gives an approximate minimizer for (OPT) and hence more efficient computation. In any event, it is clear that the feasible set S_1 of (OPT_1) can be characterized equivalently: $S_1 = \{\hat{K} \in \ell_{m \times n}^1 : K \text{ interpolates } U \text{ and } V\} = \{\hat{K} \in \ell_{m \times n}^1 : K \text{ satisfies the conditions of Lemma 1}\}$. The following results exploit the form of S_1 to establish existence first of a minimizer for (OPT_1) and then for (OPT) .

Theorem 4. Given Assumption 2, the problem (OPT_1) has a minimizer \hat{K}_0 .

Proof. For ease of notation we will give the proof only for the case that all $z_0 \in Z_{UV}$ are real and indicate how to generalize to the complex case in a straightforward way. We will show that there exists a subspace M of $\ell_{m \times n}^0$ such that $M^\perp = S_1$. Then S_1 is weak*-closed and the result follows by Corollary 1. We will use the second characterization of S_1 given above, which imposes a condition on K for each i, j, k , and z_0 . Correspondingly, we define for each condition a sequence \hat{G}_{ijkz_0} as follows:

$$\hat{G}_{ijkz_0}(l) := \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \hat{\alpha}_i^T(q-p) \hat{\beta}_j^T(p-l) [z^q]^{(k)}(z_0).$$

It is easily verified that the above sequence lies in $\ell_{m \times n}^0$ since every $z_0 \in D$. Also, given $\hat{K} \in \ell_{m \times n}^1$:

$$\begin{aligned} \langle \hat{G}_{ijkz_0}, \hat{K} \rangle &= \sum_{r=1}^m \sum_{s=1}^n \sum_{l=0}^{\infty} \\ &\quad \times \left[\sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \hat{\alpha}_i^T(q-p) \hat{\beta}_j^T(p-l) [z^q]^{(k)}(z_0) \right] \hat{K}_{rs}(l) \end{aligned}$$

$$\begin{aligned} &= \sum_{q=0}^{\infty} [\hat{\alpha}_i * \hat{K} * \hat{\beta}_j](q) [z^q]^{(k)}(z_0) \\ &= (\alpha, K\beta_j)^{(k)}(z_0). \end{aligned}$$

Let M denote the linear span of all the \hat{G}_{ijkz_0} s, a subspace of $\ell_{m \times n}^0$. Then it is immediate that $M^\perp = S_1$. For the complex zeros case, we can treat the conjugate zeros in pairs to obtain two similar sequences. \square

Theorem 5. Given Assumption 2, the problem (OPT) has a minimizer K_0 . Moreover $\Phi_0 = H - K_0$ is a polynomial transfer matrix.

Proof. To establish this we will consider the problem (OPT_1) to be posed in the primal space $\ell_{m \times n}^1$ and use the fact that S_1 , which lies in general in $\ell_{m \times n}^0$, is in fact exactly the M given in the proof of Theorem 4 and hence lies in $\ell_{m \times n}^0$. We will use the alignment condition of Theorem 1 to show first that for any minimizer \hat{K}_0 of (OPT_1) , Φ_0 has at least one row which is polynomial. Next we show that, given any minimizer of (OPT_1) for which l rows of the corresponding Φ are polynomial (where $l < m$), there exists another minimizer for which at least $l+1$ rows of the corresponding Φ are polynomial. Hence there is at least one minimizer \hat{K}_0 for which all rows of Φ_0 are polynomial. For such a \hat{K}_0 , K_0 is clearly rational and hence a minimizer for (OPT) .

First, then, suppose \hat{K}_0 is any minimizer for (OPT_1) and \hat{G}_0 is any maximizer for the dual problem $\max_{\hat{G}, \mu, \nu} \langle \hat{H}, \hat{G} \rangle$. By Theorem 1, \hat{G}_0 and $\hat{\Phi}_0$ are aligned. If \hat{G}_0 is the zero functional, then $\mu_1 = 0$ and $K_0 = H$ is a minimizer for (OPT) for which Φ_0 is a polynomial. If \hat{G}_0 is non-zero, then there is a row, say the i th, such that $\max_j \|\hat{G}_{0ij}\|_1 > 0$. Since $\hat{G}_0 \in \ell_{m \times n}^0$, there exists N such that $|\hat{G}_{0ij}(k)| < \max_j \|\hat{G}_{0ij}\|_1$ for each j and all $k > N$. It then follows easily from the alignment of $\hat{\Phi}_0$ and \hat{G}_0 that $\hat{\Phi}_{0ij}(k) = 0$ for each j when $k > N$ and hence the i -th row of Φ_0 is a polynomial.

Next, suppose $l < m$ and \hat{K}_0 is any minimizer for (OPT_1) such that l rows (say the first l) of Φ are polynomial. Partition after the l th row so that:

$$\Phi_0 = \begin{pmatrix} \hat{\Phi}_{01} \\ \hat{\Phi}_{02} \end{pmatrix} \quad \hat{H} = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix} \quad \hat{K}_0 = \begin{pmatrix} K_{01} \\ \hat{K}_{02} \end{pmatrix}$$

and consider the problem:

$$\inf_{\hat{R} \in S_2} \|(\hat{H}_2 - \hat{K}_{02}) - \hat{R}\|_1$$

where $S_2 := \{\hat{R} \in \ell_{(m-l) \times n}^1 : [\hat{K}_{01}^T (\hat{K}_{02} + \hat{R})^T]^T \in S_1\}$. Clearly if \hat{R}_0 is any minimizer for this problem, $[\hat{K}_{01}^T (\hat{K}_{02} + \hat{R}_0)^T]^T$ is a minimizer for

(OPT₁). Moreover, applying Lemma 1, we see that $\hat{R} \in S_2$ if and only if for all $z_0 \in \mathcal{D}_{UV}$, $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ we have:

$$(\alpha_i^T R \beta_j)^{(k)}(z_0) = 0, \\ k = 0, \dots, \sigma_{U_i}(z_0) + \sigma_{V_j}(z_0) - 1$$

where α_i^T denotes the last $m - l$ entries of α_i . Thus it is easily shown (cf. proof of Theorem 4) that S_2 is in c^0 and the same argument given above applies to establish the existence of a minimizer \hat{R}_0 such that $H_2 - K_{02} - R_0$ has at least one polynomial row. Hence $[\hat{K}_{01}^T (\hat{K}_{02} + \hat{R}_0)^T]^T$ is a minimizer for (OPT₁) with at least $l + 1$ polynomial rows. \square

The bad rank case

In this case, U has full column rank $= n_u$ and V has full row rank $= n_v$. We will need the following assumption:

Assumption 3. There exist n_u rows of U and n_v columns of V which are linearly independent for all z on the unit circle.

Note that this is slightly stronger than requiring that U and V have no transmission zeros on the unit circle. Under this assumption, U and V can be written in the following form without loss of generality (possibly requiring the interchange of inputs and/or outputs):

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \\ V = (\bar{V}_1 \quad \bar{V}_2)$$

where \bar{U} has dimensions $n_u \times n_u$ and is invertible and \bar{V} has dimensions $n_v \times n_v$ and is invertible. Moreover, \bar{U} and \bar{V} have no zeros on the unit circle. Thus $K = UQV$ can be written:

$$K = \begin{pmatrix} \bar{U} \\ U_2 \end{pmatrix} Q (\bar{V}_1 \quad \bar{V}_2) = \begin{pmatrix} \bar{K} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$$

and \bar{U} and \bar{V} define a good rank sub-problem of the overall problem satisfying Assumption 2. Also, we can define polynomial coprime factorizations as follows:

$$U_2 \bar{U}^{-1} = \bar{D}_U^{-1} \bar{N}_U \\ \bar{V}^{-1} \bar{V}_2 = \bar{N}_V \bar{D}_V^{-1} \tag{3.3}$$

Using these definitions we state the following result characterizing the feasible set S_1 for this case.

Theorem 6. Given U and V as above, Assumption 3, and $K \in A$, there exists $Q \in A$ satisfying $K = UQV$ if and only if:

$$(i) \begin{pmatrix} -\bar{N}_U & \bar{D}_U \end{pmatrix} \begin{pmatrix} \bar{K} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} = 0$$

$$(ii) \begin{pmatrix} \bar{K} & K_{12} \end{pmatrix} \begin{pmatrix} -N_V \\ D_V \end{pmatrix} = 0$$

(iii) \bar{K} interpolates \bar{U} and \bar{V} .

Proof. (\Rightarrow): If $K = UQV$ then certainly $\bar{K} = \bar{U}Q\bar{V}$. Now, \bar{U} and \bar{V} have good rank and satisfy Assumption 2 so that if $Q \in A$ then, by Theorem 3, (iii) holds. Also, \bar{U} and \bar{V} are invertible so that $Q = \bar{U}^{-1} \bar{K} \bar{V}^{-1}$ and hence:

$$K_{12} = \bar{U}QV_2 \Rightarrow K_{12} = \bar{K} \bar{V}^{-1} V_2 \\ K_{21} = U_2 Q \bar{V} \Rightarrow K_{21} = U_2 \bar{U}^{-1} \bar{K} \tag{*}$$

$$((*) \text{ and } K_{22} = U_2 Q V_2) \\ \Rightarrow U_2 \bar{U}^{-1} \bar{K} \bar{V}^{-1} V_2 = U_2 \bar{U}^{-1} K_{12}.$$

Using the polynomial coprime factorizations (3.3), we obtain (i) and (ii).

(\Leftarrow): Since \bar{U} and \bar{V} are invertible, there always exists a unique $Q := \bar{U}^{-1} \bar{K} \bar{V}^{-1}$ solving $\bar{K} = \bar{U} Q \bar{V}$. Since \bar{U} and \bar{V} have good rank and satisfy Assumption 2, Theorem 3 holds and (iii) $\Rightarrow Q \in A$. Also:

$$(ii) \Rightarrow K_{12} = \bar{K} \bar{V}^{-1} V_2 = \bar{U} Q V_2 \\ (i) \Rightarrow K_{21} = U_2 \bar{U}^{-1} \bar{K} = U_2 Q \bar{V} \tag{**}$$

$$((i) \text{ and } (**)) \Rightarrow K_{22} = U_2 \bar{U}^{-1} K_{12} = U_2 Q V_2$$

so that $K = UQV$.

Remark 1. If conditions (i) and (ii) of Theorem 6 are satisfied, it is straightforward to verify that also:

$$(K_{21} \quad K_{22}) \begin{pmatrix} -N_V \\ D_V \end{pmatrix} = 0.$$

This remark will prove useful in the next section. The following theorem uses the above characterization of S_1 to establish existence of a minimizer for (OPT₁).

Theorem 7. Given U and V as above and Assumption 3, the problem (OPT₁) has a minimizer \hat{K}_0 .

Proof. For each of the conditions (i), (ii) and (iii) of Theorem 6 define a subspace of all $\hat{K} \in \mathcal{C}_{n_v \times n_u}^1$ which satisfy the corresponding condition; $S_{(i)}$ for condition (i), and so on. Then $S_1 = S_{(i)} \cap S_{(ii)} \cap S_{(iii)}$. We will show that each of these subspaces and hence S_1 is weak*-closed, and the result follows. First note that $S_{(iii)}$ can be handled exactly as S_1 was in Theorem 4, by constructing a c^0 sequence corresponding to each condition in Lemma 1. The only difference is that only \bar{K} will be required to interpolate \bar{U} and \bar{V} so that the sequences will have the following form, where the partitioning is conformal with

our usual partition of K :

$$\hat{G}_{ijkz_0} := \begin{pmatrix} \tilde{G}_{ijkz_0} & 0 \\ 0 & 0 \end{pmatrix}.$$

The annihilator M^\perp of the linear span M of these sequences will be equal to $S_{(u)}$ and hence $S_{(u)}$ is *weak**-closed.

Considering now $S_{(v)}$, and letting $T := [-\tilde{N}_U \ \tilde{D}_U]$ for notational convenience, define a bounded linear operator $F: \ell^1_{n_1 \times n_u} \mapsto \ell^1_{(n_1 - n_u) \times n_u}$ as follows, given $\hat{K} \in \ell^1_{n_1 \times n_u}$:

$$F\hat{K} := \hat{T} * \hat{K}.$$

Then $S_{(v)} = \mathcal{N}(F)$, where $\mathcal{N}(\cdot)$ denotes the null space. Now define a bounded linear operator $G: \ell^0_{(n_1 - n_u) \times n_u} \mapsto \ell^0_{n_1 \times n_u}$ as follows, given $\hat{\alpha} \in \ell^0_{(n_1 - n_u) \times n_u}$:

$$(G\hat{\alpha})_{ij}(k) = \sum_{l=0}^{\infty} \sum_{m=1}^{n_1 - n_u} \hat{T}_{im}^l(l-k) \hat{\alpha}_{mj}(l) \\ i = 1, \dots, n_1 \\ j = 1, \dots, n_u \\ k = 0, 1, \dots$$

Then $F = G^*$, the adjoint of G , since, for all $\hat{\alpha} \in \ell^0_{(n_1 - n_u) \times n_u}$ and $\hat{K} \in \ell^1_{n_1 \times n_u}$:

$$\langle G\hat{\alpha}, \hat{K} \rangle = \sum_{i=1}^{n_1} \sum_{j=1}^{n_u} \sum_{k=0}^{\infty} \hat{K}_{ij}(k) \\ \sum_{l=0}^{\infty} \sum_{m=1}^{n_1 - n_u} \hat{T}_{im}^l(l-k) \hat{\alpha}_{mj}(l) \\ = \sum_{m=1}^{n_1 - n_u} \sum_{j=1}^{n_u} \sum_{l=0}^{\infty} \hat{\alpha}_{mj}(l) \\ \times \left[\sum_{i=1}^{n_1} \sum_{k=0}^{\infty} \hat{T}_{mi}(l-k) \hat{K}_{ij}(k) \right] \\ = \langle \hat{\alpha}, F\hat{K} \rangle.$$

Thus $S_{(v)} = \mathcal{N}(G^*)$ is *weak**-closed by Rudin (1973, Thm 4.12). It is clear by a similar argument that $S_{(u)}$ is also *weak**-closed. \square

In cases where only one of U and V has bad rank, say V , Assumption 3 is unchanged and we partition $V = [\tilde{V} \ V_{12}]$ where \tilde{V} has dimensions $n_v \times n_v$ and is invertible and $K = [\hat{K} \ K_{12}]$. Then U and \tilde{V} define a good rank sub-problem satisfying Assumption 2, and the conditions of Theorem 6 are modified; (i) disappears, (ii) is unchanged, and (iii) becomes " \hat{K} interpolates U and \tilde{V} ". Finally, existence of a minimizer can still only be established for (OPT_1) .

4. TRUNCATED PROBLEMS

In this section, we consider the problem of computing minimizers for (OPT) when they are known to exist or otherwise at least computing

approximate minimizers which are arbitrarily good (i.e. whose distance from H can be made arbitrarily close to the infimal distance μ_{opt}). We will again consider the good rank and bad rank cases separately, but first we will discuss the characterization of (OPT) as an infinite linear program and define, given a problem (OPT) , a class of related problems which we will call *truncated problems* which are equivalent to *finite* linear programs and which can be used in many cases to obtain exact or approximate minimizers for (OPT) . We will characterize exactly when this approach can be applied. We will also note that the assumptions required on unit circle zeros of U and V in the last section can be dropped when we consider the problem (OPT) directly.

Linear programs and truncated problems

In the previous section, we characterized the feasible set S_i of (OPT_i) for the good rank case in Lemma 1 and Theorem 3 and for the bad rank case in Theorem 6. The following modified versions of these results characterize the feasible set S of (OPT) and do not require the Assumptions 2 or 3. The notation is as established in the last section for their respective cases and the proofs are omitted as they are easily obtained by modifying the proofs of the corresponding results.

Lemma 3 (Good rank). Given $K \in \mathcal{HA}$, there exists $Q \in \mathcal{HA}$ satisfying $K = UQV$ if and only if for all $z_0 \in \mathcal{Z}_{UV}$, $i \in \{1, \dots, n_1\}$ and $j \in \{1, \dots, n_u\}$ we have:

$$(\alpha_i K \beta_j)^{(k)}(z_0) = 0, \\ k = 0, \dots, o_{U_i}(z_0) + o_{V_j}(z_0) - 1.$$

Theorem 8 (Good rank). Given $K \in \mathcal{HA}$, there exists $Q \in \mathcal{HA}$ satisfying $K = UQV$ if and only if K interpolates U and V .

Theorem 9 (Bad rank). Given $K \in \mathcal{HA}$, there exists $Q \in \mathcal{HA}$ satisfying $K = UQV$ if and only if conditions (i), (ii) and (iii) of Theorem 6 are satisfied.

Recalling that the closed loop transfer function $\Phi = H - K$ (so that $K = H - \Phi$) and using the definition of the A -norm, (OPT) can be written:

$$\inf_{\Phi \in S} \max_{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_u} |\Phi_{ij}(k)|$$

subject to:

$$H - \Phi \in S.$$

In both the good and bad rank cases, the requirement $H - \Phi \in S$ can be interpreted as a set of linear equality constraints on Φ as follows.

In the good rank case, we can use Lemma 3 above and define a set of sequences \hat{G}_{ijkz_0} corresponding to each condition as we did in the proof of Theorem 4. Then we see that $H - \Phi \in S$ if and only if for each such sequence:

$$\langle \hat{G}_{ijkz_0}, \hat{\Phi} \rangle = \langle \hat{G}_{ijkz_0}, \hat{H} \rangle.$$

This clearly defines a *finite* set of linear equality constraints on $\hat{\Phi}$. The condition of Theorem 8 that K interpolate U and V can be similarly interpreted to yield a smaller but equivalent set of constraints (i.e. the conditions of Lemma 3 contain linear dependencies not present in the conditions of Definition 2 for interpolation).

In the bad rank case, condition (iii) defines a similar finite set of linear equality constraints on $\hat{\Phi}$ (cf. proof of Theorem 7). Condition (i) is clearly satisfied if and only if:

$$[(-\hat{N}_U, \hat{D}_U) * \hat{\Phi}](k) = [(-\hat{N}_U, \hat{D}_U) * \hat{H}](k) \\ k = 0, 1, 2, \dots$$

which again defines a set of linear equality constraints on $\hat{\Phi}$ but in this case an *infinite* number. Condition (ii) defines a similar infinite set of constraints.

With these observations and the use of a standard linear programming technique for handling the objective function (see, e.g. Chvátal, 1983) we see that (OPT) is equivalent to the following linear program in the impulse response coefficients $\hat{\Phi}_{ij}(k)$ of the closed loop system and the auxiliary variables $\hat{\Phi}_{ij}^+(k)$, $\hat{\Phi}_{ij}^-(k)$, and λ :

$$(LP): \quad \inf \lambda$$

subject to:

$$\begin{aligned} \hat{\Phi}_{ij}^+(k) - \hat{\Phi}_{ij}^-(k) &= \hat{\Phi}_{ij}(k) & i &= 1, \dots, n_z \\ \hat{\Phi}_{ij}^+(k), \hat{\Phi}_{ij}^-(k) &\geq 0 & j &= 1, \dots, n_u \\ & & k &= 0, 1, \dots \end{aligned}$$

$$\sum_{j=1}^{n_s} \sum_{k=0}^{\infty} [\hat{\Phi}_{ij}^+(k) + \hat{\Phi}_{ij}^-(k)] \leq \lambda \quad i = 1, \dots, n_z$$

$$H - \Phi \in S.$$

Because (LP) has an infinite number of variables and (at least in the bad rank case) constraints, it cannot be solved directly using general linear programming techniques. Instead of (OPT), then, we study the following family of related finite dimensional problems which we call *truncated problems*, and which are indexed by the non-negative integer δ :

$$(OPT_\delta): \quad \min_{K \in S_\delta} \|H - K\|_A =: \mu_\delta$$

where $S_\delta := \{K \in S: H - K \text{ is a polynomial of degree } \leq \delta\}$. With this restriction on the degree of $\Phi = H - K$, (OPT_δ) is equivalent to the

following linear program:

$$(LP_\delta): \quad \min \lambda$$

subject to:

$$\begin{aligned} \hat{\Phi}_{ij}^+(k) - \hat{\Phi}_{ij}^-(k) &= \hat{\Phi}_{ij}(k) & i &= 1, \dots, n_z \\ \hat{\Phi}_{ij}^+(k), \hat{\Phi}_{ij}^-(k) &\geq 0 & j &= 1, \dots, n_u \\ & & k &= 0, \dots, \delta \end{aligned}$$

$$\sum_{j=1}^{n_s} \sum_{k=0}^{\delta} [\hat{\Phi}_{ij}^+(k) + \hat{\Phi}_{ij}^-(k)] \leq \lambda \quad i = 1, \dots, n_z$$

$$H - \Phi \in S_\delta.$$

For any δ , (LP_δ) has a finite number of variables and, in the good rank case, a finite number of constraints. In the bad rank case there remain in general an infinite number of constraints but we will see that in all cases of interest these are equivalent to a finite set so that each (LP_δ) that we wish to solve is a *finite* linear program.

Of course, the problem (OPT_δ) always has a bounded objective function but the set S_δ may be empty for a given problem (OPT) and a given δ . In this case μ_δ is not defined. In order to address this problem, we state the following condition.

Condition 2. There exists δ^* such that S_{δ^*} is non-empty or, equivalently, such that (OPT) has a feasible point for which $\Phi = H - K$ is a polynomial of degree δ^* .

If (and only if) this condition is satisfied we can define a monotonically increasing integer sequence $(\delta(i))_{i=0}^\infty$ for which, by taking $\delta(0) := \delta^*$, the corresponding sequence of problems (OPT_{δ(i)}) will have well defined infimal norms $\mu_{\delta(i)}$. The sequence $\mu_{\delta(i)}$ will clearly be monotonically non-increasing and, moreover, the following theorem will establish that:

$$\lim_{i \rightarrow \infty} \mu_{\delta(i)} = \mu_{\text{opt}}.$$

Theorem 9. Given Condition 2, $\mu_\delta - \mu_{\text{opt}}$ can be made arbitrarily small by taking δ sufficiently large.

Proof. Given Condition 2, there exists $K_{\delta^*} \in S_{\delta^*}$ and we can write for any $K \in S$:

$$\begin{aligned} \|H - K\|_A &= \|(H - K_{\delta^*}) + (K_{\delta^*} - K)\|_A \\ &= \|\Phi_{\delta^*} - (K - K_{\delta^*})\|_A \end{aligned}$$

where Φ_{δ^*} is a polynomial of degree $\leq \delta^*$. Moreover, $K \in S$ if and only if $(K - K_{\delta^*}) \in S$ so that:

$$\mu_{\text{opt}} = \inf_{K \in S} \|H - K\|_A = \inf_{K \in S} \|\Phi_{\delta^*} - K\|_A.$$

Thus there is a $K' = UQ'V$ for which $\|\Phi_{\delta^*} - K'\|_A$ approximates μ_{opt} arbitrarily closely. Also,

the set of polynomials is dense in $\mathcal{H}A$ and U and V are in A so that K' can be approximated arbitrarily closely by approximating Q' sufficiently closely with a polynomial Q_p . Finally, since U and V can be taken to be polynomial (Dahleh and Pearson, 1987), we have that $\Phi_p := \Phi_{\delta^*} - UQ_pV$ is a polynomial whose norm is arbitrarily close to μ_{opt} . Thus by taking δ to be the degree of Φ_p , we have the result. \square

With this theorem we know that whenever Condition 2 is satisfied, a procedure for finding arbitrarily good approximate minimizers is to simply formulate and solve a sequence of problems (LP_δ) for increasingly large δ . If we choose δ too small, the problem will be infeasible and δ must be increased but the existence of δ^* ensures that a feasible problem will eventually be obtained in this way. In the following we examine conditions under which Condition 2 is satisfied and summarize the application of this solution procedure in the good rank and bad rank cases separately.

The good rank case

If Assumption 2 is satisfied (no unit circle zeros of U or V), Condition 2 is clearly satisfied since by Theorem 5 we can take δ^* to be the degree of $H - K_0$ where K_0 is a minimizer for (OPT) . The following lemma establishes that even when Assumption 2 is not satisfied, Condition 2 still is.

Lemma 4. Given U and V with good rank, there exists δ^* such that S_{δ^*} is non-empty.

Proof. Recalling the Smith-McMillan form decompositions (3.1) of U and V , consider the related problem (OPT') defined by $H' := L_U^{-1}HR_V^{-1}$, $U' := M_U R_U$, and $V' := L_V M_V$. Then $L_U = R_V = I$ and the condition of Lemma 3 becomes, for each $z_0 \in \mathcal{I}_{U^{-1}V^{-1}} (= \mathcal{I}_{U'V'})$, $i \in \{1, \dots, n_z\}$, and $j \in \{1, \dots, n_a\}$:

$$(K'_{ij})^{(k)}(z_0) = 0 \quad k = 0, \dots, \sigma_{U'}(z_0) + \sigma_{V'}(z_0) - 1.$$

For each i and j let $\Phi'_{ij}(z)$ be a scalar polynomial such that $K'_{ij} := H'_{ij} - \Phi'_{ij}$ satisfies the above for each z_0 (such a polynomial always exists). Then there exists a $Q' \in \mathcal{H}A$ such that $\Phi' = H' - U'Q'V'$ and hence $K := UQ'V$ is a feasible point of (OPT) for which $\Phi = H - K = L_U\Phi'R_V$ is a polynomial. Let δ^* be the degree of Φ . Then $\Phi \in S_{\delta^*}$. \square

Recall that (LP_δ) is a finite linear program for any δ in the good rank case. Then the situation is as follows. By Lemma 4, it is always possible to formulate a sequence of feasible finite linear

programs corresponding to an increasing sequence of δ s. It is possible to estimate δ^* in order to aid in formulating this sequence, but in practice simply increasing δ until a feasible program is obtained is usually satisfactory. By Theorem 9, the corresponding μ_δ s converge to μ_{opt} from above. At present, we have no method which allows δ to be selected *a priori* such that $|\mu_\delta - \mu_{\text{opt}}|$ is less than a given ϵ . When no unit circle zeros are present, however, the exact minimizer corresponds to a polynomial Φ and hence can be found by solving (LP_δ) for a suitably chosen δ . This δ may be estimated using a generalization of the bound given in Dahleh and Pearson (1987) but again, in practice, simply increasing δ until the degree of the minimizing Φ stops increasing is usually satisfactory.

The bad rank case

We have seen in the good rank case that Condition 2 always holds. This is not so in the bad rank case. The following theorem characterizes for bad rank problems exactly when it is satisfied. (Here H is partitioned as K was in Theorem 6.)

Theorem 10. Given U and V with bad rank, Condition 2 is satisfied if and only if the transfer matrices T_{UH} and T_{UV} defined:

$$T_{UH} := (-\tilde{N}_U \quad -\tilde{D}_U) \begin{pmatrix} H & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$$

$$T_{UV} := \begin{pmatrix} \tilde{H} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \begin{pmatrix} -N_V \\ D_V \end{pmatrix}$$

are both polynomials.

Proof. (\Rightarrow) : If Condition 2 is satisfied, (OPT) has a feasible point K for which $\Phi = H - K$ is a polynomial. K must satisfy conditions (i) and (ii) of Theorem 6 and hence the condition of Remark 1. Since the matrices $[-\tilde{N}_U \quad -\tilde{D}_U]$ and $[-N_V^T \quad D_V^T]^T$ are polynomial, so are T_{UH} and T_{UV} . (\Leftarrow) : Recall that the matrices $[-\tilde{N}_U \quad -\tilde{D}_U]$ and $[-N_V^T \quad D_V^T]^T$ are left (right) polynomial coprime factorizations of $U_U U^{-1}(\tilde{V}^{-1}V_U)$. There exist associated right (left) polynomial coprime factorizations $U_U \tilde{U}^{-1} = N_U D_U^{-1}$ ($\tilde{V}^{-1}V_U = \tilde{D}_V^{-1}\tilde{N}_V$) and the following Bezout identities can be constructed:

$$\begin{pmatrix} \tilde{X}_U & -\tilde{Y}_U \\ -\tilde{N}_U & \tilde{D}_U \end{pmatrix} \begin{pmatrix} D_U & Y_U \\ N_U & X_U \end{pmatrix} = \begin{pmatrix} \tilde{D}_V & \tilde{N}_V \\ Y_V & \tilde{X}_V \end{pmatrix} \begin{pmatrix} X_V & -N_V \\ -Y_V & D_V \end{pmatrix}$$

where all the blocks are polynomial. Define:

$$B_U := \begin{pmatrix} \tilde{X}_U & -\tilde{Y}_U \\ -\tilde{N}_U & \tilde{D}_U \end{pmatrix} \quad B_V := \begin{pmatrix} X_V & -N_V \\ -Y_V & D_V \end{pmatrix}$$

and note that B_U^{-1} and B_V^{-1} are both polynomial. Now consider a problem (OPT') defined by taking $H' := B_U H B_V$, $U' = B_U U$, and $V' = V B_V$. Then it is straightforward to show that U' and V' have the following forms:

$$U' = \begin{pmatrix} \tilde{U}' \\ 0 \end{pmatrix} \quad V' = (\tilde{V}' \quad 0)$$

where \tilde{U}' and \tilde{V}' are square and full rank. Also, partitioning as usual, the following blocks of H' :

$$(H'_{21} \quad H'_{22}) = T_{U'U} B_V \begin{pmatrix} H'_{12} \\ H'_{22} \end{pmatrix} = B_U T_H$$

are polynomials since $T_{U'U}$ and T_{UV} are.

Now \tilde{H}' , \tilde{U}' , and \tilde{V}' define a good rank problem which, by Lemma 4, has a feasible point for which $\tilde{\Phi}' = \tilde{H}' - \tilde{K}'$ is polynomial and hence there exists $Q \in \mathcal{RA}$ such that $\tilde{\Phi}' = \tilde{H}' - \tilde{U}' Q \tilde{V}'$. Then clearly:

$$K' := \begin{pmatrix} \tilde{U}' Q \tilde{V}' & 0 \\ 0 & 0 \end{pmatrix}$$

is a feasible point of (OPT') for which $\Phi' = H' - K'$ is given by:

$$\Phi' = \begin{pmatrix} \tilde{\Phi}' & H'_{12} \\ H'_{21} & H'_{22} \end{pmatrix}$$

and is hence polynomial. Finally, then, $K = B_U^{-1} K' B_V^{-1} = U Q V$ is a feasible point of (OPT) for which $\Phi = H - U Q V = B_U^{-1} \Phi' B_V^{-1}$ is polynomial. \square

Now suppose that (OPT) satisfies the condition of the theorem and hence Condition 2 is satisfied. Then for $\delta \geq \delta^*$ the constraint equations of (LP_δ) corresponding to condition (i) of Theorem 6 are equivalent to the following finite set, where $\delta_i := \text{degree}([- \tilde{N}_i \quad \tilde{D}_i])$:

$$[[- \tilde{N}_i \quad \tilde{D}_i] * \tilde{\Phi}](k) = \tilde{T}_{U'U}(k) \\ k = 0, \dots, \delta + \delta_i$$

and similarly for condition (ii). The set of constraint equations corresponding to condition (iii) is also finite so that (LP_δ) is a finite linear program for each $\delta \geq \delta^*$.

The situation in the bad rank case is then as follows. Recalling the results of Section 3, we have only established the existence of a minimizing K for (OPT_1) in the bad rank case, not for (OPT) . In particular, there need not exist a minimizing K for (OPT) for which the corresponding Φ is a polynomial. Hence, (OPT) cannot in general be solved exactly by the solution of (LP_δ) for any finite δ . However, when the conditions of Theorem 10 are satisfied, it is possible to formulate a sequence of feasible finite linear programs corresponding to an increasing sequence of δ s. It is possible, as in the good rank case, to estimate δ^* to aid in the

formulation of this sequence if desired. As Theorem 9 shows, the solution of such a sequence of problems yields a corresponding sequence of μ_δ s which converge from above to μ_{opt} . Again as in the good rank case, we have no method which allows δ to be selected *a priori* such that $|\mu_i - \mu_{\text{opt}}|$ is less than a given ϵ .

* Finally, we remark on the case when only U or V say V , has had rank and recall that in this case we partition $V = [\tilde{V} \quad V_{12}]$. The conditions in Theorem 10 then reduce to requiring the matrix:

$$T_{UV} := (\tilde{H} \quad H_{12}) \begin{pmatrix} -N_V \\ D_V \end{pmatrix}$$

to be polynomial and, provided it is, the preceding discussion applies to the approximate solution of such problems as well.

5. CONSTRAINED PROBLEMS

Given any standard problem [say (OPT) defined by H , U , and V], two index sets \mathcal{J}_1 and \mathcal{J}_2 which partition $\{1, \dots, n_z\}$ and which have n_1 and n_2 elements, respectively, where $n_i \geq 1$, and a set of positive real numbers $\{d_i\}_{i \in \mathcal{J}_1}$, we can define an associated *constrained* ℓ^1 optimization problem. In this section we observe that the computation of approximate minimizers for almost all such problems which have a non-empty feasible set and for which (OPT) satisfies Condition 2 can be handled using a very minor modification of the method given in the last section for solving (OPT) itself. That is, arbitrarily good approximate minimizers can be found by solving a sequence of finite linear programs.

To state the constrained problem associated with (OPT) assume without loss of generality that $\mathcal{J}_1 = \{1, \dots, n_1\}$ and $\mathcal{J}_2 = \{n_1 + 1, \dots, n_z\}$ and define a weighting matrix $W_i := \text{diag}[(d_i^{-1})_{i \in \mathcal{J}_1}]$. Also define the following partitions:

$$H = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix} \quad U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \quad K = \begin{pmatrix} K_1 \\ K_2 \end{pmatrix}$$

where H_1 has n_1 rows, H_2 has n_2 rows, and so on. Then the constrained problem is:

$$(OPT_C): \quad \inf_{K \in S_C} \|H_1 - J_1\|_A =: \mu_C$$

where $S_C := \{K \in \mathcal{KA} : \exists Q \in \mathcal{RA} \text{ satisfying } K = U Q V \text{ and } \|W_i(H_i - K_i)\|_A \leq 1\}$. This problem corresponds to minimizing the maximum ℓ^1 -norm of a set of *regulated* outputs indexed by \mathcal{J}_1 , subject to the constraint that the ℓ^∞ -norms of the *constrained* outputs indexed by \mathcal{J}_2 are bounded by the corresponding d_i .

Whereas the feasible set S of (OPT) is always non-empty (since always $0 \in S$), S_C need not be

and hence μ_c may not be well defined. We can describe when S_c is non-empty in terms of the infimal norm μ_F of an associated standard problem defined by $H_F := W_c H_c$, $U_F := W_c U_c$ and $V_F := V$. Clearly $S_c \neq \emptyset$ if and only if $\mu_F \leq 1$. In order to apply the approach of the last section to computing approximate solutions to feasible problems (*OPTC*), we define truncated constrained problems exactly as we did for unconstrained:

$$(OPTC_\delta): \quad \inf_{K \in S_{c,\delta}} \|H_r - K_r\|_A =: \mu_{c,\delta}$$

where $S_{c,\delta} := \{K \in S_c: H - K \text{ is polynomial of degree } \leq \delta\}$.

If (*OPT*) satisfies Condition 2 and also $\mu_r < 1$, it can be shown using essentially the ideas of the proof of Theorem 9 that (*OPTC* _{δ}) has a feasible point for sufficiently large δ and moreover that the sequence of infimal norms $\mu_{c,\delta}$ for an increasing sequence of δ 's has as its limit μ_c .

Finally, we observe that when (*OPTC* _{δ}) has a feasible point it is equivalent to the following finite linear program:

$$(LP_\delta): \quad \min \lambda$$

subject to:

$$\begin{aligned} \hat{\Phi}_n^+(k) - \hat{\Phi}_n^-(k) &= \hat{\Phi}_n(k) & i &= 1, \dots, n_r \\ \hat{\Phi}_n^+(k), \hat{\Phi}_n^-(k) &\geq 0 & j &= 1, \dots, n_w \\ & & k &= 0, \dots, \delta \end{aligned}$$

$$\sum_{k=0}^{\delta} \sum_{i=1}^{n_r} [\hat{\Phi}_n^+(k) + \hat{\Phi}_n^-(k)] \leq 2\lambda \quad i \in \mathcal{I}$$

$$\sum_{k=0}^{\delta} \sum_{i \in \mathcal{J}} [\hat{\Phi}_n^+(k) + \hat{\Phi}_n^-(k)] \leq d_i \quad i \in \mathcal{J},$$

$$H - \Phi \in S_\delta$$

where S_δ is the feasible set of (*OPT* _{δ}).

It is interesting to note that while such constrained problems are known to be difficult when other norms such as the H^2 -norm are considered (Ting and Poolla, 1988), they are handled extremely simply when using the A -norm. In fact, the complexity of (*LPC* _{δ}) is no greater than that of the corresponding unconstrained problem.

6. CONCLUSION

In this paper, we have extended the results of Dahleh and Pearson (1987, 1988) concerning existence of minimizers for ℓ^1 -optimal control problems to a broader class of plants and provided more direct proofs. The only assumption required is to preclude open loop poles or zeros on the unit circle. A class of truncated

problems corresponding to polynomial closed loop transfer matrices and which are equivalent to finite linear programs is defined. Necessary and sufficient conditions are given for when a sequence of such problems can be solved to obtain either exact or approximate minimizers. While we have at present no method for determining in general the size of approximation error introduced by this procedure, this problem has been addressed in a special case in (Staffans, 1990). In that paper, the scalar mixed sensitivity problem studied in Dahleh and Pearson (1988) is considered and a procedure is given for determining the approximation error as closely as desired. In addition, a minimizer for which the closed loop transfer matrix is rational (i.e. non-polynomial) is found for the same problem.

We have also observed that problems incorporating norm constraints on some outputs while regulating others can be solved at least approximately using only a slight modification of the method for solving unconstrained problems. The simplicity of constrained problems is a property of the ℓ^1 -norm which sets it apart from the H^2 -norm and other norms frequently used for control system design.

Acknowledgement This research was supported by NASA under Grant NAG9-208 and by the National Science Foundation under Grants ECS-8806977 and CCR-8809615.

REFERENCES

- Cheng, L. and J. B. Pearson (1978) Frequency domain synthesis of multivariable linear regulators. *IEEE Trans. Aut. Control*, **AC-23**, 3-15.
- Churchill, R. V. and J. W. Brown (1984) *Complex Variables and Applications*. McGraw-Hill, New York.
- Chvátal, V. (1983) *Linear Programming*. Freeman, New York.
- Dahleh, M. A. and J. B. Pearson (1987) ℓ^1 Optimal feedback controllers for MIMO discrete time systems. *IEEE Trans. Aut. Control*, **AC-32**, 314-322.
- Dahleh, M. A. and J. B. Pearson (1988) Optimal rejection of persistent disturbances, robust stability, and mixed sensitivity minimization. *IEEE Trans. Aut. Control*, **AC-33**, 722-731.
- Francis, B. A. (1987) *A Course in H_2 Control Theory*. Springer, New York.
- Luenberger, D. G. (1969) *Optimization by Vector Space Methods*. Wiley, Chichester, U.K.
- MacFarlane, A. G. J. and N. Karamias (1976) Poles and zeros of linear multivariable systems: a survey of the algebraic, geometric and complex variable theory. *Int. J. Control*, **24**, 33-74.
- Rudin, W. (1973) *Functional Analysis*. McGraw-Hill, New York.
- Staffans, O. J. (1990) The mixed sensitivity minimization problem has a rational ℓ^1 -optimal solution. Helsinki University of Technology Institute of Mathematics Research Reports, A274.
- Ting, T. and K. Poolla (1988) Upper bounds and approximate solutions for multidisk problems. *IEEE Trans. Aut. Control* **AC-33**, 783-786.
- Vidyasagar, M. (1987) Further results on the optimal rejection of persistent bounded disturbances, Part I: The discrete-time case. preprint.

Minimal Structure in the Block Decoupling Problem with Stability*

C. COMMAULT†‡, J. M. DION† and J. A. TORRES†‡

New feedback invariants permit characterization of the simplest block decoupled system achievable for a given linear system using linear stabilizing control laws.

Key Words—Decoupling; stability; multivariable control systems; transfer function analysis; algebraic system theory.

Abstract—In this paper we consider the block decoupling problem with stability. We characterize the minimal McMillan degree achievable for the block decoupled system. For this, we introduce a new list of integers which are feedback invariant. These integers are related in a simple way to the minimal number of unstable and infinite zeros of the block decoupled system.

1. INTRODUCTION

IN THE recent years a great deal of interest has been devoted to the structural features of decoupling problems. Various lists of integers strongly related with the deep structure of the system have been introduced. These integers have been successfully used for characterizing both the existence of specific decoupling control laws and the simplest achievable structure for the decoupled system.

In particular:

In Descusse and Dion (1982), the classical static state feedback decoupling problem is revisited. The solvability condition of Falb and Wolovich (1967) can be expressed as follows: the system infinite structure is given by the infinite structure of the transfer matrix rows.

In Commault *et al.* (1986), the essential orders characterizing the minimal infinite structure of the decoupled system are presented for the general row by row decoupling problem.

In Descusse *et al.* (1988), structural conditions are given for the solvability of the row by row decoupling problem by static state feedback (with possible singular input transformation).

* Received 24 April 1989; revised 1 December 1989; received in final form 18 July 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. F. Curtain under the direction of Editor H. Kwakernaak.

† Laboratoire d'Automatique de Grenoble, ENSIEG-INPG-CNRS, B.P. 46-38402 Saint-Martin-d'Hères, France.

‡ Present address: CINVESTAV del IPN, Apartado Postal 14-740, Mexico 07000 D.F., Mexico

§ Author to whom all correspondence should be addressed.

In Dion and Commault (1988), the dynamic state feedback decoupling problem is considered. The minimal infinite and unstable structure of the decoupled system is characterized.

In Hautus and Heyman (1983), necessary and sufficient conditions for solving block decoupling problems with stability are given. In Dion *et al.* (1990), the minimal infinite structure of the block decoupled system is characterized.

In this paper, we consider the block decoupling problem with stability. We characterize the minimal McMillan degree achievable for the block decoupled system incorporating stability requirements. For this we introduce a new list of integers which are feedback invariant. Moreover we give the minimal infinite and unstable structure of the i th block of the decoupled system. This structure is easily obtained by comparing the unstable and infinite structure of the original system transfer matrix (stabilized if necessary by an adequate static state feedback) and the one of the transfer matrix obtained by removing the i th row block.

The paper is organized as follows. In Section 2 the block decoupling problem with stability is formulated, and some notations and preliminaries are also given. Two useful technical results concerning Hermite forms are stated.

Section 3 is devoted to the study of a new list of integers $\{n_{i,c}\}$ which is invariant by stability preserving static state feedback. In Section 4 we prove that the minimal McMillan degree achievable for the decoupled system with stability is v the sum of the $n_{i,c}$.

In Section 5 an illustrative example exhibits the main features of the proposed approach.

2. PROBLEM FORMULATION AND PRELIMINARIES

Notation

Let $\mathbf{R}(s)$ be the field of rational functions and $\mathbf{R}[s]$ the ring of polynomials. A rational function

$f(s) = n(s)/d(s)$ is said to be proper (resp. strictly proper) if $\deg(d(s)) \geq \deg(n(s))$ (resp. $\deg(d(s)) > \deg(n(s))$) where $\deg(n(s))$ denotes the polynomial degree of $n(s)$. The function $f(s) = n(s)/d(s)$ will be called a proper rational stable function if it is proper and if all the roots of $d(s)$ are stable. The stability domain under consideration is any region in the complex plane symmetrically located with respect to the real axis and including at least one point of the real axis.

Denote by $\mathbf{R}_p(s)$ the ring of proper rational functions and $\mathbf{R}_p^{p \times m}(s)$ the set of proper rational $p \times m$ transfer matrices. The units (invertible elements) of the ring $\mathbf{R}_p^{m \times m}(s)$ are called bicausal matrices and are characterized by the property that $B(s)$ is a bicausal matrix if and only if

$$\det \left(\lim_{s \rightarrow \infty} B(s) \right) \neq 0.$$

The set of proper rational stable functions, denoted $\mathbf{R}_{ps}(s)$, is an Euclidean domain with degree function $\gamma(f(s)) = \gamma_1(f(s)) + \gamma_2(f(s))$, where $\gamma_1(f(s))$ is the infinite zero order of $f(s)$ and $\gamma_2(f(s))$ is the number of unstable zeros of $f(s)$ counted with their multiplicity (Hung and Anderson, 1979; Callier and Desoer, 1982).

The set of $p \times m$ matrices with entries in $\mathbf{R}_{ps}(s)$ will be denoted by $\mathbf{R}_{ps}^{p \times m}(s)$. The units of $\mathbf{R}_{ps}^{m \times m}(s)$ are called bicausal-bistable matrices and are characterized by the property that $B(s)$ is bicausal-bistable if and only if $\det B(s)$ is a unit of $\mathbf{R}_{ps}(s)$.

Problem formulation

Introduce now the block decoupling problem with stability.

We will consider now the stable block decoupling problem. In order to avoid trivialities we will require the compensated system to be just as "output controllable" as the origin system is.

We will say that the proper precompensator $C(s)$ is *admissible* if

$$\text{rang } T(s)C(s) = \text{rang } T(s).$$

This admissibility condition is equivalent to the preservation of the C^* controlled output trajectories (see Brockett and Mesarovic, 1965).

Let $T(s)$ be a $p \times m$ proper rational matrix, partitioned in row-blocks relatively to a given list of positive integers (p_1, \dots, p_k) , such that $\sum_{i=1}^k p_i = p$, in the following way:

$$T(s) = \begin{bmatrix} T_1(s) \\ \vdots \\ T_k(s) \end{bmatrix}$$

with $T_i(s) \in \mathbf{R}_p^{p_i \times m}(s)$ for $i = 1, \dots, k$.

The system with transfer matrix $T(s)$ is said to

be block decoupled relatively to the partition $\{p_i\}$ if there exist positive integers m_1, \dots, m_k satisfying $\sum_{i=1}^k m_i = m$, such that $T(s)$ has the block diagonal form:

$$T(s) = \begin{bmatrix} T_{11}(s) & & 0 \\ & \ddots & \\ 0 & & T_{kk}(s) \end{bmatrix} = \text{diag} (T_{11}(s), \dots, T_{kk}(s))$$

with $T_{ii}(s) \in \mathbf{R}_p^{p_i \times m_i}(s)$ for $i = 1, \dots, k$.

This means that each above defined input block influences only one output block. If one wants this influence to be effective the $T_{ii}(s)$ must be non-null for each i , in this case the system is called nondegenerate.

We will consider here the decoupling problem with stability, i.e. the compensated decoupled system is stable and no internal pole zero cancellation occurs.

In order to (block) decouple a given system (Σ)

$$\begin{aligned} \dot{x} &= Ax + Bu & u &\in \mathbf{R}^m \\ x &\in \mathbf{R}^n \\ y &= Cx & y &\in \mathbf{R}^p \\ T(s) &= C(sI - A)^{-1}B \end{aligned}$$

we will use a *combined compensator* of the form:

$$u = Fx + C(s)v$$

where x is the state of a minimal realization of (Σ) , $F \in \mathbf{R}^{m \times n}$, and $C(s)$ is a proper stable precompensator not necessarily square.

This is a quite general form of compensation. Both parts of the control law are necessary because:

- if one uses pure precompensation $F = 0$ and if (Σ) is unstable, unstable pole zero cancellations will occur between $C(s)$ and $T(s)$.
- Conditions for static or dynamic state feedback decoupling are very restrictive, (Descusse *et al.* 1988; Dion and Commault, 1988).

On the other hand Hautus and Heymann (1983) proved that (Σ) is decouplable by precompensation if and only if:

$$\text{rang } T(s) = \sum_{i=1}^k \text{rang } T_i(s).$$

Because of stability requirements $C(s)$ will necessarily be stable.

The stable block decoupling problem can be formulated as follows:

Let $T(s)$ be a $p \times m$ proper rational matrix partitioned in row-blocks relatively to (p_1, \dots, p_k) . Is it possible to exhibit an admissible combined compensator such that the compensated system is block decoupled, non

degenerate, stable with no unstable pole zero cancellations?

As shown in the following lemma, there is no loss of generality in assuming that the system is stable.

Lemma 1. Consider the system Σ and the closed loop system Σ_F where F is any state feedback such that $A + BF$ has stable eigenvalues. The above defined stable block decoupling problem is solvable by combined compensation if and only if stable block decoupling by pure precompensation is achievable on Σ_F .

Proof. Sufficiency: clearly a pure precompensation on Σ_F is a combined compensation on Σ .

Necessity: Let $u = F_1x + C(s)v$ be a combined stabilizing block decoupling compensator on Σ . Then $u = (F_1 - F)x + C(s)v$ is a combined stabilizing block decoupling compensator on Σ_F which is equivalent to the pure block decoupling precompensator $B_1(s)C(s)$ where

$$\begin{aligned} B_1(s) &= [I - (F_1 - F)(sI - A - BF)^{-1}B]^{-1} \\ &= [I + (F_1 - F)(sI - A - BF_1)^{-1}B] \end{aligned}$$

acting on Σ_F .

$B_1(s)$ is then bicausal and bistable. Both $C(s)$ and $B_1(s)C(s)$ are stable. $B_1(s)C(s)$ is therefore a stable block decoupling pure precompensator acting on Σ_F . \square

In the sequel, we will then assume that the system under consideration is stable. We look for stable block decoupling pure precompensators.

In this paper we focus our attention on the simplest achievable decoupled systems with stability. More precisely we will give the minimal McMillan degree and the minimal infinite and unstable structure achievable for the decoupled system ensuring the stability of the compensated system.

Preliminaries

In what follows we use the symbol $\delta_M(T(s))$ to denote the McMillan degree of a transfer matrix $T(s)$. By McMillan degree we mean the polynomial degrees sum of the denominators in the classical Smith-McMillan form over $\mathbf{R}[s]$ of the transfer matrix $T(s)$ (Kailath, 1980).

In the sequel, we will need Smith-McMillan forms of proper rational transfer matrices over the Euclidean ring of proper stable rational functions $\mathbf{R}_p(s)$. Let us introduce such factorizations, which were studied in Verghese (1978) and Vardulakis and Karcianas (1983).

Definition 1. Let $T(s)$ be a $p \times m$ rational matrix of rank r , a Smith-McMillan factorization of $T(s)$ over $\mathbf{R}_p(s)$ is a factorization of the form:

$$T(s) = B_{11}(s)\Lambda(s)B_{12}(s)$$

where $B_{11}(s)$ and $B_{12}(s)$ are bicausal-bistable

matrices and

$$\Lambda(s) = \begin{bmatrix} \Delta(s) & 0 \\ 0 & 0 \end{bmatrix}$$

where $\Delta(s) = \text{diag}(n_1(s)/d_1(s), \dots, n_r(s)/d_r(s))$, with

$$n_i(s) = \frac{\epsilon_i(s)}{\pi^i} \quad \text{and} \quad d_i(s) = \psi_i(s)$$

$n_i(s)$ and $d_i(s)$ are coprime in $\mathbf{R}_p(s)$ (i.e. $\epsilon_i(s)$, $\psi_i(s)$ have no common unstable zeros and $n_i(s)$ or $d_i(s)$ is bicausal). $\epsilon_i(s)$, $\psi_i(s)$ are coprime polynomials which have no stable roots and $\pi = s + a$ is an arbitrary stable polynomial. Furthermore $\epsilon_i(s)$ divides $\epsilon_{i+1}(s)$ and $\psi_{i+1}(s)$ divides $\psi_i(s)$. The integers t_i (resp. f_i) are decreasingly (resp. increasingly) ordered.

Consider the rational transfer matrix, the stability being considered in the usual sense:

$$T(s) = \begin{bmatrix} \frac{1}{s+1} & \frac{1}{s-3} \\ \frac{1}{s+5} & \frac{1}{s-7} \end{bmatrix}$$

it can be shown that it has two infinite zeros of orders (1, 2), two unstable poles at $s = 3, 7$ and one unstable zero at $s = 1$. The Smith-McMillan form of $T(s)$ over $\mathbf{R}_p(s)$ is:

$$\Delta(s) = \begin{bmatrix} 1/\pi & 0 \\ 0 & \frac{(s-1)/\pi^2}{(s-3)(s-7)/\pi^2} \end{bmatrix}$$

where $d_1(s) = 1$ and $d_2(s)$ is bicausal

which exhibits the above given unstable and infinite structure of $T(s)$.

Theorem 1. Let $T(s)$ be a $p \times m$ rational matrix. Then there exist Smith-McMillan factorizations of $T(s)$ over $\mathbf{R}_p(s)$. Moreover $\Lambda(s)$ is uniquely defined modulo $\pi = s + a$, a stable polynomial, and is called the Smith-McMillan form of $T(s)$ over $\mathbf{R}_p(s)$. The unstable zeros (resp. poles) of $T(s)$ are the roots of $\epsilon_i(s)$ (resp. $\psi_i(s)$) and the infinite zero (resp. pole) orders of $T(s)$ are given by $f_i = \deg(\epsilon_i(s))$ (resp. $t_i = \deg(\psi_i(s))$).

Let $d(T(s))$ denote the sum of number of zeros at infinity plus number of unstable zeros plus number of poles at infinity plus number of unstable poles, multiplicities included:

$$d(T(s)) = \sum (f_i + t_i). \quad (2.1)$$

For a proper stable rational matrix, which has no unstable and infinite poles ($\psi_i(s) = 1$ and $t_i = 0$, for $i = 1, \dots, k$) $d(T(s)) = \sum_{i=1}^r f_i$. In this case the Smith-McMillan factorization of $T(s)$ over $\mathbf{R}_p(s)$ coincides with the usual Smith factorization over the ring of proper stable rational functions.

Recall now the Hermite form over $\mathbf{R}_p(s)$, the ring of proper rational stable functions.

Theorem 2 (Morse, 1975). Let $T(s)$ be a $r \times m$ proper rational stable matrix of rank r . $T(s)$ can be factorized in $T(s) = H(s)B(s)$, where $H(s) \in \mathbf{R}_p^{r \times m}(s)$ and $B(s)$ is a bicausal-bistable matrix (unit of $\mathbf{R}_p^{m \times m}(s)$) and $H(s) = [\tilde{H}(s) \ 0]$, with $\tilde{H}(s) \in \mathbf{R}_p^{r \times r}(s)$:

$$\tilde{H}(s) = \begin{bmatrix} \delta_1/\pi^n & & 0 \\ & \ddots & \\ h_{ij}(s) & & \delta_r/\pi^n \end{bmatrix}$$

where $\pi = s + u$ is an arbitrary stable polynomial of degree one and

$$h_{ij} = \frac{\gamma_{ij}}{\pi^{n_{ij}}}$$

$n_{ij} < n_i$ with n_{ij}, n_i positive integers and δ_i, γ_{ij} polynomials, δ_i possessing no stable zeros. $H(s)$ is called a Hermite form of $T(s)$ over $\mathbf{R}_p(s)$.

$H(s)$ is nonunique, and is defined up to units of $\mathbf{R}_p(s)$. Uniqueness of $H(s)$ may be obtained by adding some conditions (Morse, 1975).

We will give now two technical lemmas concerning the Hermite forms over $\mathbf{R}_p(s)$.

Since $B(s)$ is a bicausal-bistable matrix, the infinite and unstable zeros of $H(s)$ are those of $T(s)$. Let α_i denote the maximal infinite zero order of $H(s)$ and let q be the number of different finite unstable zeros z_1, \dots, z_q of $H(s)$ and denote by α_i the maximal order of the finite zero z_i . Define $n := \alpha_i + \alpha_z$, we have:

Lemma 2. Let $T(s)$ be a $r \times m$ proper rational stable matrix of rank r and $H(s)$ be a Hermite form of $T(s)$ over $\mathbf{R}_p(s)$. With the above notations the McMillan degree of each entry of $H(s)$ is less than or equal to n .

Proof. Note that the zero orders of $H(s)$ are the same as those of $\tilde{H}(s)$ and the pole orders of $\tilde{H}^{-1}(s)$ are the same as the zero orders of $\tilde{H}(s)$.

The maximal pole order of a rational matrix at any given point is an upper bound for the maximal pole orders of all entries at the same point. Then n is an upper bound for the infinite plus unstable zero orders of $\tilde{H}^{-1}(s)$ and in particular n is an upper bound for the infinite plus unstable zero orders of the diagonal elements of $\tilde{H}(s)$, i.e. $n_{ii} \leq n$.

Therefore the McMillan degree n_{ij} of $h_{ij}(s)$ is lower than or equal to n , due to the known Hermite form property $n_{ij} \leq n_i$. ∇

Lemma 3. Let $T(s)$ be a $r \times m$ proper rational stable matrix of rank r . Factorize $T(s)$ as follows: $T(s) = H(s)B(s)$, where $H(s) = [\tilde{H}(s) \ 0]$ is a Hermite form of $T(s)$ over $\mathbf{R}_p(s)$ as in Theorem 2. With the previous notations we

have the following:

$$d(T(s)) = d(H(s)) = \sum_{i=1}^r n_i.$$

Proof. First, observe that the infinite and unstable zero structure of $T(s)$ is equal to that of $H(s)$, since $B(s)$ is a bicausal-bistable matrix.

$$d(T(s)) = d(H(s)) = d(\tilde{H}(s)). \quad (2.2)$$

Let

$$T(s) = B_1(s) \begin{bmatrix} \Delta(s) & 0 \end{bmatrix} B_2(s),$$

$$\Delta(s) = \text{diag}(\epsilon_1(s)/\pi_1', \dots, \epsilon_r(s)/\pi_r'),$$

be a Smith-McMillan form over $\mathbf{R}_p(s)$, $\epsilon_i(s)$ has no common factors with $\pi_j'(i, j = 1, \dots, r)$, then:

$$d(\tilde{H}(s)) = d(\det \Delta(s)).$$

There exists a Smith-McMillan factorization of $\tilde{H}(s)$ over $\mathbf{R}_p(s)$, $\tilde{H}(s) = B_1'(s) \Delta'(s) B_2'(s)$, with $\Delta'(s) = \Delta(s)$, such that:

$$\det \tilde{H}(s) = b_1(s) \det(\Delta(s)) b_2(s)$$

with $b_1(s), b_2(s)$ bicausal-bistable rational functions, which implies that:

$$d(\det \tilde{H}(s)) = d(\det \Delta(s))$$

the last expression allows us to write:

$$d(\tilde{H}(s)) = d(\det \tilde{H}(s))$$

and from the particular structure of $\tilde{H}(s)$ the result follows.

3. FEEDBACK INVARIANTS

In this section we will consider a $p \times m$ strictly proper rational transfer matrix $T(s)$ of rank r , partitioned in row-blocks relatively to a given list of positive integers (p_1, \dots, p_k) , such that $\sum_{i=1}^k p_i = p$, in the following way:

$$T(s) = \begin{bmatrix} T_1(s) \\ \vdots \\ T_k(s) \end{bmatrix}$$

with $T_i(s) \in \mathbf{R}_p^{p_i \times m}(s)$ for $i = 1, \dots, k$.

Let $T'(s)$ denote the matrix obtained from $T(s)$ by removing its i th row-block, i.e.

$$T'(s) = \begin{bmatrix} T_1(s) \\ \vdots \\ T_{i-1}(s) \\ T_{i+1}(s) \\ \vdots \\ T_k(s) \end{bmatrix}$$

In what follows the following lemma (Dion *et al.* 1990) will be useful.

Lemma 4. Let $T(s)$ be a $p \times m$ proper rational matrix partitioned in row-blocks relatively to (p_1, \dots, p_k) , such that $\sum_{i=1}^k p_i = p$. Let r_i denote the rank of $T_i(s)$, for $i = 1, \dots, k$.

If $T(s)$ is such that:

$$\text{rank } T(s) = \sum_{i=1}^k r_i$$

then $T(s)$ can be decomposed as:

$$T(s) = U(s)\tilde{I}\tilde{T}(s)$$

where $\tilde{T}(s)$ is a $r \times m$ full row rank proper rational matrix partitioned in row-blocks relatively to $\{r_i\}$. $U(s) = \text{diag}(U_1(s), \dots, U_k(s))$, $U_i(s)$ is a $p_i \times p_i$ unimodular matrix and $\tilde{I} = \text{diag}(\tilde{I}_1, \dots, \tilde{I}_k)$, $\tilde{I}_i = \begin{bmatrix} I_{r_i} \\ 0 \end{bmatrix}$; I_{r_i} is the $r_i \times r_i$ identity matrix.

We will define now a new list of integers $n_{i,cs}$, $i = 1, \dots, k$, and show that these integers are feedback invariant. These invariants generalize the block decoupling invariants introduced in Dion *et al.* (1990), when introducing stability requirements.

Definition 2. Let $T(s)$ be a $p \times m$ strictly proper stable rational matrix decomposed as in Lemma 4 according to the output partition (p_1, \dots, p_k) , i.e. $T(s) = U(s)\tilde{I}\tilde{T}(s)$ and let $\tilde{T}(s)$ be factorized as:

$$\tilde{T}(s) = [R(s) \ 0]B_i(s)$$

where $R(s)$ is a $r \times r$ proper rational nonsingular matrix and $B_i(s)$ is a bicausal-bistable matrix. Consider $R^{-1}(s)$ partitioned as follows:

$$R^{-1}(s) = [\tilde{R}_1(s), \dots, \tilde{R}_k(s)]; \quad \tilde{R}_i(s) \in \mathbf{R}^{r \times r_i}(s).$$

We will define the stable block decoupling invariants, denoted $n_{i,cs}$, of $T(s)$ as the total number of infinite and unstable poles of $\tilde{R}_i(s)$ counted with their multiplicity. Since $\tilde{R}(s)$ possess no infinite and unstable zeros, one has:

$$n_{i,cs} = d(\tilde{R}_i(s)) \quad \text{for } i = 1, \dots, k$$

and we call $v = \sum_{i=1}^k n_{i,cs}$ the stable block decoupling degree of $T(s)$.

From the above definition v and $n_{i,cs}$ are invariant under stability preserving static state feedback on a minimal realization of $T(s)$. This comes from the fact that $T(s)$ being stable, such feedbacks are equivalent to bicausal-bistable precompensators.

We will prove later that the $n_{i,cs}$ are independent of the chosen factorization.

The following lemma will be used later to provide us with a nice characterization of $n_{i,cs}$.

Lemma 5. Let $G(s)$ be a $r \times r$ full rank strictly proper stable rational matrix, partitioned in row-blocks relatively to (r_1, \dots, r_k) . Partition $G^{-1}(s)$, the inverse matrix of $G(s)$, in column-blocks relatively to (r_1, \dots, r_k) . Let $\tilde{G}_i(s)$ denote the i th column-block of $G^{-1}(s)$, then:

$$d(\tilde{G}_i(s)) = d(G(s)) - d(G'(s)) \quad \text{for } i = 1, \dots, k$$

where $G'(s)$ denotes the matrix obtained from $G(s)$ by removing the i th row-block.

The lemma can be proved following closely the proof of Lemma 2 given in Dion *et al.* (1990). For this, $\mathbf{R}_p(s)$ will be replaced by $\mathbf{R}_{ps}(s)$, $d_+(.)$ will be replaced by $d(.)$. *Mutatis mutandis* the result follows, using Lemma 3 of the present paper.

Theorem 3. Let $T(s)$ be a $p \times m$ strictly proper stable rational matrix decomposed in row-blocks relatively to (p_1, \dots, p_k) , such that $\sum_{i=1}^k p_i = p$. Let $T' = \sum_{i=1}^k r_i$, where $r = \text{rank } T(s)$ and r_i

$\text{rank } T_i(s)$. Let $T'(s)$ be the matrix obtained from $T(s)$ by removing the i th row-block. Then the stable block decoupling invariants $n_{i,cs}$ of $T(s)$ satisfy:

$$n_{i,cs} = d(T(s)) - d(T'(s)) + \sigma(T_i(s))$$

$$\text{for } i = 1, \dots, k$$

where $\sigma(T_i(s))$ denotes the sum of the degrees of a minimal polynomial basis for the left kernel of $T_i(s)$.

Proof. Begin by decomposing $T(s)$ as in Lemma 4:

$$T(s) = U(s)\tilde{I}\tilde{T}(s) \quad (3.1)$$

and denote $\tilde{T}(s) = [R(s) \ 0]B_i(s)$ a Hermite factorization of $\tilde{T}(s)$ over $\mathbf{R}_{ps}(s)$, where $B_i(s)$ is a bicausal-bistable matrix. So, by definition:

$$n_{i,cs} = d(\tilde{R}_i(s)) \quad \text{for } i = 1, \dots, k$$

where $\tilde{R}_i(s)$ denotes the i th column block of $R^{-1}(s)$.

From Lemma 5:

$$n_{i,cs} = d(R(s)) - d(R'(s)) \quad \text{for } i = 1, \dots, k$$

by definition of $d(R(s))$ (see expression 2.1) and since $B_i(s)$ is a bicausal-bistable matrix we have:

$$n_{i,cs} = d(\tilde{T}(s)) - d(\tilde{T}'(s)) \quad \text{for } i = 1, \dots, k. \quad (3.2)$$

In the next part, we prove that $d(\tilde{T}(s)) = d(T(s)) + \sigma(T(s))$, which will be also true for $T'(s)$, i.e. $d(\tilde{T}'(s)) = d(T'(s)) + \sigma(T'(s))$. In this way, by the row-block independence hypothesis ($\text{rank } T(s) = \sum r_i$) we will have $\sigma(T(s)) - \sigma(T'(s)) = \sigma(T_i(s))$, which proves the theorem.

For this, observe that the right null space of $\tilde{T}(s)$, denoted $\text{Ker } \tilde{T}(s)$, is equal to $\text{Ker } T(s)$. In fact, $\text{Ket } \tilde{T}(s)$, is a subspace of $\text{Ker } T(s)$ and $\dim(\text{Ker } \tilde{T}(s)) = \dim(\text{Ker } T(s)) = m - r$.

Denote $\sigma_R(T(s))$ the sum of the degrees of a minimal polynomial basis for $\text{Ker } T(s)$, so we have:

$$\sigma_R(T(s)) = \sigma_R(\tilde{T}(s)) \quad (3.3)$$

which is also true for $T'(s)$, i.e. $\sigma_R(T'(s)) = \sigma_R(\tilde{T}'(s))$.

Now, define $Z(T(s))$ as the number of finite stable zeros of $T(s)$ counted with their

multiplicity. Since $U(s)$ of (3.1) is unimodular, we can write:

$$Z(T(s)) = Z(\tilde{T}(s)). \quad (3.4)$$

Moreover, we have that the McMillan degree of $\tilde{T}(s)$ is equal to the McMillan degree of $T(s)$, which allows us to write:

$$\begin{aligned} \sigma(T(s)) + d(T(s)) + Z(T(s)) + \sigma_R(T(s)) \\ = d(\tilde{T}(s)) + Z(\tilde{T}(s)) + \sigma_R(\tilde{T}(s)) \end{aligned}$$

and from (3.3) and (3.4), we obtain:

$$\sigma(T(s)) + d(T(s)) = d(\tilde{T}(s))$$

and similarly: $\sigma(T'(s)) + d(T'(s)) = d(\tilde{T}'(s))$. This combined with (3.2) lead us to:

$$n_{i,\sigma} = d(T(s)) - d(T'(s)) + \sigma(T(s)) - \sigma(T'(s))$$

which ends the proof, since $\sigma(T(s)) - \sigma(T'(s)) = \sigma(T_i(s))$, as noted above. ∇

Remark. From this theorem it is clear that the $n_{i,\sigma}$ do not depend on the chosen factorization in Definition 2.

4. THE MINIMAL BLOCK DECOUPLING PROBLEM WITH STABILITY

In this section we look for the minimal McMillan degree achievable for decoupled system ensuring the stability of the compensated system. We will show that the stable block decoupling invariants $n_{i,\sigma}$ defined in the previous section, represent such a minimal degree for the i th diagonal block of the decoupled system

As shown in Lemma 1, there is no loss of generality in assuming that the system is stable.

Let us give a lower bound for the McMillan degree of the decoupled system.

Lemma 6. Let $T(s)$ be a $p \times m$ proper stable rational matrix of rank r , partitioned in row-blocks relatively to (p_1, \dots, p_k) and let $C(s)$ be a $m \times r$ proper stable admissible precompensator which block decouples $T(s)$ with stability relatively to (p_1, \dots, p_k) . The McMillan degree of the i th block of the compensated system $D_i(s)$ is greater than or equal to $n_{i,\sigma}$, where $T(s)C(s) = \text{diag}(D_1(s), \dots, D_k(s))$.

The proof follows closely the proof for the case without stability requirements (see Dion *et al.* (1990)). It was given by contradiction, assuming that there exists a block decoupling precompensator such that the McMillan degree of $D_i(s)$ is lower than $n_{i,\sigma}$, which leads to a nonproper or nonstable precompensator.

We are now ready to give the main result of this section:

Theorem 4. Let $T(s)$ be a $p \times m$ proper stable rational matrix of rank r , partitioned in row-blocks relatively to (p_1, \dots, p_k) , such that

$\sum_{i=1}^k p_i = p$. If the system $T(s)$ can be block decoupled relatively to (p_1, \dots, p_k) by an admissible proper precompensator with stability, then the minimal McMillan degree achievable for the block decoupled system is v , the stable block decoupling degree of $T(s)$. In this case the i th diagonal block of the decoupled system has McMillan degree $n_{i,\sigma}$.

When particularized to the row by row decoupling problem of surjective systems this result is given in Dion and Commault (1988).

Proof. By Lemma 6 the minimal McMillan degree achievable for the decoupled system is greater than or equal to v . We will construct a block decoupling stable precompensator and then we will show that each block of the decoupled system has McMillan degree $n_{i,\sigma}$.

First, decompose $T(s)$ as in Lemma 4:

$$T(s) = U(s)\tilde{T}(s)$$

and consider a factorization of $\tilde{T}(s)$ over $\mathbf{R}_p(s)$ of the form:

$$\tilde{T}(s) = [R(s) \ 0]B_i(s)$$

where $B_i(s)$ is a bicausal-bistable matrix and $R(s)$ is a full rank proper stable rational matrix. A possible solution is to consider a Hermite form of $T(s)$ over $\mathbf{R}_p(s)$.

Now denote $\tilde{R}_i(s)$ the i th column-block of $R^{-1}(s)$. Define β_i , the maximal infinite pole order of $\tilde{R}_i(s)$. Let q_i denote the number of different unstable poles a_{i1}, \dots, a_{iq_i} of $\tilde{R}_i(s)$ and denote β_{ij} the maximal order of the pole a_{ij} , for $j = 1, \dots, q_i$ and $i = 1, \dots, k$. Define the integers g_i as follows:

$$g_i := \beta_i + \sum_{j=1}^{q_i} \beta_{ij} \quad \text{for } i = 1, \dots, k$$

and also define

$$G_i(s) := \frac{p_i(s)}{\pi^{g_i}} \tilde{R}_i(s) \quad \text{for } i = 1, \dots, k \quad (4.1)$$

with $p_i(s) = \prod_{j=1}^{q_i} (s - a_{ij})^{\beta_{ij}}$ for $i = 1, \dots, k$ and $\pi = s + a$ is a stable monomial. $G_i(s)$ is a proper stable rational matrix.

Consider now a Hermite form $[L_i^T(s) \ 0]$ of $G_i^T(s)$ over $\mathbf{R}_p(s)$, where T denotes transposition, one has:

$$G_i(s) = B_i(s) \begin{pmatrix} L_i(s) \\ 0 \end{pmatrix} \quad \text{for } i = 1, \dots, k \quad (4.2)$$

where $B_i(s)$ is a bicausal-bistable matrix and $L_i(s)$ is a full rank proper stable rational matrix.

We are now able to give the following stable compensator:

$$C(s) = B_i^{-1}(s) \begin{bmatrix} Y(s) \\ X(s) \end{bmatrix} \quad (4.3)$$

where $Y(s) = R^{-1}(s) \text{diag} (D_1(s), \dots, D_k(s))$

$$D_i(s) := L_i^{-1}(s) \frac{p_i(s)}{\pi_i^k} \quad \text{for } i = 1, \dots, k$$

and $X(s)$ is any $(m-r) \times r$ proper stable rational matrix. Note that $Y(s)$ can be written as follows:

$$Y(s) = [Y_1(s), \dots, Y_k(s)];$$

using (4.1), (4.2) one has:

$$Y_i(s) = G_i(s)L_i^{-1}(s) = B_i(s) \begin{bmatrix} L_i \\ 0 \end{bmatrix}$$

then $C(s)$ is a proper stable compensator.

Recall that β_{ix} and β_{iy} denote the maximal orders of the infinite and unstable poles of $\hat{R}_i(s)$. From (4.1) and using a Smith-McMillan factorization of $\hat{R}_i(s)$ over $\mathbf{R}_{pi}(s)$ it follows that the maximal orders of the infinite (resp. unstable) zeros of $G_i(s)$ are lower than or equal to β_{ix} (resp. β_{iy}).

$L_i(s)$ is a Hermite form; using Lemma 2 it follows that each entry of $L_i(s)$ has a denominator degree lower than or equal to g_i . It follows that:

$$D_i^{-1}(s) = L_i(s) \frac{\pi_i^k}{p_i(s)}.$$

Then $D_i^{-1}(s)$ has no stable poles, so $D_i(s)$ has no stable zeros. The McMillan degree $\delta_M(D_i(s))$ of the nonsingular matrix $D_i(s)$ is then equal to $d(D_i(s))$ the sum of the infinite and unstable zeros counted with their multiplicity:

$$\delta_M(D_i(s)) = d(D_i(s)).$$

From the definition 2.1

$$d(D_i(s)) = d(D_i^{-1}(s)).$$

Using (4.1), (4.2) and the definition of $n_{i, \infty}$ we get:

$$d(D_i^{-1}(s)) = d\left(L_i(s) \frac{\pi_i^k}{p_i(s)}\right) = d(\hat{R}_i(s)) = n_{i, \infty}.$$

Therefore

$$\delta_M \text{diag} (D_1(s) \cdots D_k(s)) = v.$$

Recall that

$$T(s)C(s) = U(s)\hat{I} \text{diag} (D_1(s), \dots, D_k(s)).$$

Since $U(s)$ is unimodular and due to the particular structure of \hat{I} we finally get:

$$\delta_M(T(s)C(s)) = v.$$

5. AN ILLUSTRATIVE EXAMPLE

Consider the system:

$$T(s) = \begin{bmatrix} \frac{1}{(s+1)^2} & 0 & 0 & \frac{1}{(s+1)^4} \\ 0 & \frac{1}{(s+1)^2} & 0 & \frac{1}{(s+1)^4} \\ \frac{1}{s+1} & \frac{1}{s+1} & \frac{s-1}{(s+1)^3} & \frac{s+2}{(s+1)^3} \end{bmatrix}.$$

Let us factorize $T(s)$ as follows: $T(s) = [R(s) \ 0]B_i(s)$ with:

$$R(s) = \begin{bmatrix} \frac{1}{(s+1)^2} & 0 & 0 \\ 0 & \frac{1}{(s+1)^2} & 0 \\ \frac{1}{s+1} & \frac{1}{s+1} & \frac{s-1}{(s+1)^3} \end{bmatrix}$$

and

$$B_i^{-1}(s) = \begin{bmatrix} (s+1)^2 & 0 & 0 \\ 0 & (s+1)^2 & 0 \\ -(s+1)^4 & -(s+1)^4 & (s+1)^3 \\ s-1 & s-1 & s-1 \end{bmatrix}.$$

The system $T(s)$ is not row by row decouplable by dynamic state feedback, since $m = 4 < 2p - k = 6 - 1$, where k is the column rank at infinity of $R^{-1}(s)$ (Dion and Commault, 1988). However, it is block decouplable by dynamic state feedback with the output partition $(2, 1)$, since $m \geq 2r - k^*$, with $m = 4$ (number of inputs), $r = \text{rank } T(s) = 3$ and the integer $k^* = 2$ (the dimension of the maximal space spanned by the column blocks of $R^{-1}(s)$ when s goes to infinity) defined in Commault *et al.* (1990).

Write $R^{-1}(s) = [\hat{R}_1(s), \hat{R}_2(s)]$, according to the partition $(2, 1)$. The matrix $\hat{R}_1(s)$ has two infinite poles of orders $(2, 3)$ and one unstable pole at $s = 1$. Matrix $\hat{R}_2(s)$ has one infinite pole of order 2, and one unstable pole at $s = 1$. So by definition the stable block decoupling invariants are $n_{1, \infty} = 6$ and $n_{2, \infty} = 3$, giving a stable block decoupling degree $v = 9$.

Since $T(s)$ is proper and stable, in this case $d(T(s))$ is the sum of the infinite and unstable zero orders of $T(s)$, counted with their multiplicity. $T(s)$ has three infinite zeros of orders $(1, 2, 3)$ and one unstable zero at $s = 1$. The matrix $T_1(s)$ has two infinite zeros of orders $(2, 2)$ and the matrix $T_2(s)$ has one infinite zero of order 1. In this example $\sigma(T_1(s)) = \sigma(T_2(s)) = 0$, then we have:

$$n_{1, \infty} = d(T(s)) - d(T_2(s)) = 7 - 1$$

$$n_{2, \infty} = d(T(s)) - d(T_1(s)) = 7 - 4$$

as stated in Theorem 3.

We will illustrate now the compensator construction given in the proof of Theorem 4. For this consider $G_i(s) = ((s-1)/(s+1)^4)\hat{R}_1(s)$, as in (4.1). Consider now a left Hermite form over $\mathbf{R}_{pi}(s)$ of $G_i(s)$:

$$G_i(s) = B_i(s) \begin{bmatrix} L_i(s) \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{s-1}{(s+1)^2} & -1 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 0 & \frac{s-1}{(s+1)^2} \\ 1 & 0 \end{bmatrix}$$

the procedure is similar for $\bar{R}_2(s)$.

Using (4.3) we get the following compensator:

$$C(s) = B_r^{-1}(s) \begin{bmatrix} Y(s) \\ X(s) \end{bmatrix};$$

with

$$Y(s) = [Y_1(s), Y_2(s)]; Y_i(s) = B_i(s) I_r,$$

for $i = 1, 2$, where $X(s)$ is any $(m-r) \times r$ proper stable rational matrix. In this way we obtain:

$$Y(s) = \begin{bmatrix} \frac{(s-1)}{(s+1)^2} & -1 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

$$Y(s) = \begin{bmatrix} \frac{(s-1)}{(s+1)^2} & -1 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

The decoupled system is in this case:

$$T(s)C(s) = \begin{bmatrix} \frac{s+1}{(s+1)^4} & -\frac{s-1}{(s+1)^2} & 0 \\ 0 & \frac{s-1}{(s+1)^2} & 0 \\ 0 & 0 & \frac{s-1}{(s+1)^4} \end{bmatrix}$$

whose McMillan degree is equal to the stable block decoupling degree $v = n_{1,rs} + n_{2,rs} = 9$, which represents the minimal McMillan degree achievable for the decoupled system.

6. CONCLUSION

In this paper we showed the importance of new feedback invariants for characterizing the minimal structure achievable for block decoupled systems with stability. The following points can also be mentioned:

From Theorem 3, $d(T(s) - d(T'(s)))$ represents the minimal number of unstable and infinite zeros achievable for the i th block of the decoupled system. In Dion *et al.* (1990), the minimal infinite structure of the decoupled system is characterized, then combining these two results the minimal unstable zero structure achievable for the decoupled system is obtained.

In Commiault *et al.* (1990) it is proven that a system can be block decoupled with stability by dynamic state feedback if the number of inputs is sufficiently large. When this condition is met the minimal infinite and unstable structure is also given by $d(T(s)) - d(T'(s))$.

In Dion *et al.* (1991) the complete infinite and unstable minimal structure for the block decoupled system with stability is detailed.

REFERENCES

- Brockett, R. and M. D. Mesarovic (1965). The reproducibility of multivariable systems *J. Math. Anal. Appl.* **11**, 548.
- Callier, F. M. and C. A. Desoer (1982). Multivariable feedback systems. Springer, New York.
- Commiault, C., J. M. Dion and J. Torres (1990). Invariant spaces of linear systems; application to block decoupling. *IEEE Trans. Aut. Control*, **AC-35**, 618-623.
- Commiault, C., J. Descusse, J. M. Dion, J. F. Lafay and M. Malabre (1986). New decoupling invariants: the essential orders. *Int. J. Control*, **44**, 689-700.
- Descusse, J. and J. M. Dion (1982). On the structure at infinity of linear square decouplable systems. *IEEE Trans. Aut. Control*, **AC-27**, 971-974.
- Descusse, J., J. F. Lafay and M. Malabre (1988). Solution to Morgan's problem. *IEEE Trans. Aut. Control*, **AC-33**, 732-739.
- Dion, J. M. and C. Commiault (1988). The minimal delay decoupling problem: feedback implementation with stability. *SIAM J. Control*, **26**, 66-82.
- Dion, J. M., C. Commiault and J. A. Torres (1991). Stable block decoupling invariants, geometric and transfer matrix characterizations. *Proc. Joint Conf. on New Trends in System Theory*, Genova. Birkhauser, Boston.
- Dion, J. M., J. A. Torres and C. Commiault (1990). New feedback invariants and the block decoupling problem. *Int. J. Control*, **51**, 219-235.
- Falb, P. L. and W. A. Wolovich (1967). Decoupling in the design and synthesis of multivariable control systems. *IEEE Trans. Aut. Control*, **AC-12**, 651-669.
- Hautus, M. L. J. and M. Heymann (1983). Linear feedback decoupling—transfer function analysis. *IEEE Trans. Aut. Control*, **AC-28**, 823-832.
- Hung, N. D. and B. D. O. Anderson (1979). Triangularization technique for the design of multivariable control systems. *IEEE Trans. Aut. Control*, **AC-24**, 455-460.
- Kailath, T. (1980). *Linear Systems*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Morse, A. S. (1975). System invariants under feedback and cascade control. *Proc. Int. Symp.* Uline Springer, Berlin.
- Vardulakis, A. I. G. and N. Karcarnias (1983). Structure, Smith-McMillan form and coprime MFDs of a rational matrix inside a region $P - \Omega \cup \{\infty\}$. *Int. J. Control*, **38**, 927-957.
- Verghese, G. (1978). Infinite frequency behaviour in generalized dynamical systems. Ph.D. Dissertation, Stanford University, CA.

Estimation of the Control Period for Self-tuners*

MIROSLAV KÁRNY†

A Bayesian estimation methodology is exploited for estimating the optimal control period leading to a well-justified and feasible procedure for determining this key characteristic of self-tuners.

Key Words—Control period; structure determination; self-adjusting systems; Bayesian identification.

Abstract—The control period is a key tuning knob of all existing discrete-time self-tuners. An algorithm for its systematic data-based choice is presented in the paper. It relies on the theory of Bayesian structure determination applied to a special class of control-rate dependent regression models. The models describe the entire measured-data history even if the inputs vary with another rate than that attributed to the identified model. At the same time, if the input signal changes with a specific rate the corresponding model reduces to the standard SISO regression.

The proposed algorithm modifies, in accordance with the observed data, prior probabilities assigned to the compared control periods. In a single formula, it weights the average predictive ability of the model, the intersampling behaviour of the output, the number of unknown model parameters and the number of data; and, the uncertainty of the parameters. Promising properties of the algorithm are illustrated by simulation results.

1. INTRODUCTION

CONVERSIONS *continuous signals* \leftrightarrow *sampled data* form inevitable parts of any digital controller manipulating a continuous-time process. The two basic rates determine the conversions:

Sampling rate: the rate with which an A/D convertor transforms the measured data into a digital form; and

Control rate: the rate with which a feedback digital signal is fed into a D/A convertor.

Assuming (as usual) that both rates are constant, we speak about sampling and control period, respectively. The adopted interpretation implies that the sampling period is always less or equal than the control period.

* Received 12 June 1989; revised 18 December 1989; received in final form 29 May 1990. The original version of this paper was presented at the 11th IFAC World Congress on Automatic Control at the Service of Mankind which was held in Tallinn, Estonia, USSR during August, 1990. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor P. J. Gawthrop under the direction of Editor P. C. Parks.

† Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, Pod vodárenskou věží 4, 18208V, Prague 8, Czechoslovakia.

The sampling period determines the rate at which information is fed into the controller and it should be as short as possible. Its lower bound is determined by technical conditions.

The control period is a key user's knob for influencing the behaviour of any digital controller. Its choice is addressed repeatedly: new rules and tools for its tuning are proposed and the impact on the controller performance studied. Nevertheless, the rule of thumb—make from 7 to 30 changes of input per rise time—dominates (at least in self-tuning control).

The paper presents an algorithm which estimates the proper control period using data collected on the process to be controlled.

The solution is found under the assumption that only non-zero sampling and consequently control periods are feasible. The assumption is technically sound and simplifies the mathematics used.

If the shortest sampling period is assigned the value 1, then the choice of the control period is equivalent to the selection of an integer n which determines:

Control conditions

- (A.) Inputs are constant‡ within n (elementary) consecutive steps;
- (B.) A single representant only of each n -tuple of outputs measured within the control period is used in the feedback.

Careful modelling is the crucial point of the solved problem. The selected class of system models have to—in addition to the Control conditions—respect:

Estimation conditions

- (A.) Inputs may change in any (elementary) step;

‡ Zero-order holder is assumed for simplicity. An extension to other cases is straightforward.

(B_r) No data should be omitted for an efficient control-period estimation.

In the paper, control-rate dependent models meeting both pairs of conditions are described in probabilistic terms. For comparison of their descriptive abilities, the theory of Bayesian structure determination (Kárný, 1983; Kárný and Kulhavý, 1988; Peterka and Kárný, 1979; Peterka, 1981) is adopted.

The case of regression models—which form a core of a wide class of self-tuners—is elaborated in detail. The off-line version is treated here. The approach is also applicable in real time if an increased computational burden is acceptable.

The layout of the paper is as follows. After introducing basic notions (Section 2), a short discussion of control-rate determination under mismodelling (Section 3) follows.

Probabilistic modelling of the closed loop is recalled in Section 4. The introduced system model makes it possible to specify a proper class of control-rate dependent models in probabilistic terms (Section 5, the main methodological result). Then Bayesian estimation and structure determination which are basic tools for solving the formulated task are reviewed in Section 6.

Control-rate dependent regression models fulfilling the requirements stated in Section 5 are proposed in Section 7. Bayesian identification is applied to these models in Sections 8 and 9 (the main practical result). Properties of the final algorithm are illustrated by a simulation example presented in Section 10.

2. NOTIONS, NOTATIONS AND CONVENTIONS

Subtleties of continuous-time random processes are avoided by assuming that the controlled system works in discrete time (τ). The time scale is determined by the sampling period which has been assigned the value 1.

Under the above agreement, the control period of a digital controller becomes an integer number $n \geq 1$ which restricts the rate of the input changes. The input can be modified by measured data at most after each n th elementary sampling time instant.

A set with a generic point x is denoted x^* ($x^* = \{x\}$).

A mapping f is often described by its value at a generic point [$f(x)$ means $f: x^* \rightarrow \dots$].

A visualization of causality is achieved in this way.

A single input $u(\cdot)$, single output $y(\cdot)$ stochastic system is assumed.

System input, system output and input-output (data) pairs at a sampling moment τ are denoted $u(\tau)$, $y(\tau)$ and $d(\tau)$, respectively.

The estimation phase is the time interval

within which data are collected for estimation only. The time interval $(T, 0]$ ($-\infty < T < 0$) is assigned to it.

The control phase is the time interval within which adaptive control takes place. The time interval $[1, T]$ ($T < \infty$) is assigned to it.

The following shorthand notation is used for time-dependent sequences defined on a sub-interval $[t, \bar{t}]$ of the estimation and control phases ($T < t \leq \bar{t} \leq T$)

$$x_t' = \{x(\tau)\}_{\tau=t-1}^t, \quad x' = x_{T+1}' = \{x(\tau)\}_{T+1}^T.$$

According to the convention made, $x_t'^*$ denotes the set $\{x_t'\}$.

Input-output behaviour of a controlled system, on a time interval of interest $T+1, \dots, 0, 1, \dots, T$, is modelled by a sequence of causal, generally stochastic, mappings

$$S(\tau): (d^{(\tau-1)*}, u^*(\tau)) = (y^{(\tau-1)*}, u^*(\tau)) \rightarrow y^*(\tau). \quad (1)$$

$S^T \in S^{T*}$ is called system model.

A controller is described by a sequence of causal, generally stochastic, mappings

$$R(\tau): d^{(\tau-1)*} = (y^{(\tau-1)*}, u^{(\tau-1)*}) \rightarrow u^*(\tau). \quad (2)$$

$R^T \in R^{T*}$ is called the control strategy. Notice that the obligatory step of the transport delay is attached to the controller. This convention is made in order to respect agreements applied in the majority of references cited here.

The symbols denoting the system model and the control strategy are further simplified

$$S = S^T, \quad R = R^T.$$

The symbols $p(\cdot)$ and $p(\cdot | \cdot)$, $r(\cdot | \cdot)$, $s(\cdot | \cdot)$ are reserved for probability density functions (defined with respect to a suitable dominating measure) and conditional probability density functions, respectively. The term (conditional) probability density function is abbreviated to (c.)p.d.f. The c.p.d.f.'s $r(\cdot | \cdot)$, $s(\cdot | \cdot)$ are reserved for descriptions of the control strategy and of the system model, respectively. Otherwise $p(\cdot)$ or $p(\cdot | \cdot)$ are used.

Various (c.)p.d.f.'s are distinguished by their arguments.

3. MISMODELLING AND CONTROL PERIOD

Let S be a fixed system model and R_0 be the optimal strategy defined as a minimizing argument

$$R_0 \in \text{Arg min}_{R \in R^*} \mathcal{K}(R, S)$$

where $\mathcal{K}(\cdot, \cdot)$ denotes a non-negative functional which introduces complete ordering on a set of admissible control strategies R^* .

Let R_0^* denote a subset of admissible control strategies. Clearly, it holds

$$\min_{R \in R_0^* \subset R^*} \mathcal{K}(R, S) \geq \min_{R \in R^*} \mathcal{K}(R, S). \quad (3)$$

R_n^* is interpreted as the set of controllers with control period n . Inequality (3) and the observation

$$n_1 \leq n_2 \Rightarrow R_{n_2}^* \subset R_{n_1}^*$$

imply: the smaller control period, the better control quality can be achieved. This formal observation contradicts engineering experience.

The contradiction is usually faced by judging whether much quality can be gained by shortening the control period. A more realistic setting should take into account a mismodelling influence. It means that the "true" system description *need not* belong to the set of models assumed.

Let S_n denote the system model used for choosing the strategy $R_n \in R_n^*$, i.e. having control period n . The optimal strategy R_{opt} is selected within R_n^* as a minimizing argument

$$R_{opt} \in \text{Arg min}_{R_n \in R_n^*} \mathcal{K}(R_n, S_n).$$

The performance of this strategy is to be judged according to the value of the functional $\mathcal{K}(R_{opt}, S_n)$, where $S_n \neq S_0$ denotes the "true" controlled system.

Clearly, the strategy R_{opt} can be outperformed by a strategy related to a longer control period because the discrepancy of $\mathcal{K}(R_{opt}, S_n)$ and $\mathcal{K}(R_{opt}, S_0)$ increases for fast control.

4 INFORMATION NEEDED FOR CONTROL DESIGN UNDER UNCERTAINTY

It is widely accepted that a complete ordering of admissible strategies—which is used for choosing the optimal strategy—has to face both randomness of the controlled process and incompleteness of the system description.

It follows from the above inspection that such a complete ordering can rely only on an approximate description of the controlled system.

Detailed discussion of the suitable theory is beyond the scope of this paper. We claim, however, that the control design based on general model of statistical decision making with a Bayesian view on incompletely known (random, uncertain) entities (De Groot, 1970) fulfils the requirements. The mismodelling (approximation) aspects of Bayesian decision making can be understood from the paper (Harris and Heinel, 1979).

For stochastic systems, the expected value $\mathcal{E}[\cdot]$ of a non-negative loss function $L(\cdot)$

$$L: d^{T*} = (u^{T*}, y^{T*}) \rightarrow [0, \infty] \quad (4)$$

is taken as the strategy-ordering criterion, i.e.

$$\mathcal{K}(R, S) = \mathcal{E}[L(d^T)] \quad d^T \in d^{T*}. \quad (5)$$

The loss function (4) is formally defined on the union of the estimation and control phases for sake of notational simplicity. Of course, the

strategy optimization is performed for the control phase only.

We restrict ourselves to the common case when the expectation can be expressed in terms of a probability density function. Under this assumption, the closed loop description [given by the mappings (1), (2)] is specified if for any generic data d^T the c.p.d.f. $p(d^T)$ is given.

Now, let us distinguish in $p(d^T)$ the probabilistic description of the mappings (1), (2).

The following basic properties of the conditional expectation $\mathcal{E}[\cdot | \cdot]$ (Rao, 1987) will be used:

for any conditions a, b , it holds

$$\mathcal{E}[\cdot | a] = \mathcal{E}[\mathcal{E}[\cdot | a, b] | a]; \quad (6)$$

if the condition in the conditional expectation is generated by observed random variables, the conditional expectation can be expressed as a function of them.

Using the property (6), the criterion (5) can be evaluated for a given causal strategy (2) by the backward recursion

$$\begin{aligned} W(d^{T-1}) &= \mathcal{E}[\mathcal{E}[W(d^T) | d^{T-1}, u(T)] | d^{T-1}], \\ \tau &= T, T-1, T-2, \dots \end{aligned} \quad (7)$$

$$W(d^T) = L(d^T), \quad \mathcal{K}(R, S) = W(d^T).$$

Inspecting the recursion (7) with the second property of the conditional expectation in mind, we see that a causal control strategy has to specify (and is specified by) the collection of the c.p.d.f.'s

$$R = \{r(u(\tau) | d^{T-1}) : u(\tau) \in u^*(\tau), d^{T-1} \in d^{T-1*}, \tau \in (T, T]\}. \quad (8)$$

From the right-hand side of (7), it can also be seen that the system model and the strategy description have to specify the c.p.d.f.'s $p(y(\tau), d'_{\tau+1} | d^{T-1}, u(\tau))$ for $\tau = T+1, \dots, T$. The description of the control strategy (8) and the chain rule for c.p.d.f.'s imply that the system model is given by the set of c.p.d.f.'s

$$S = \{s(y(\tau) | d^{T-1}, u(\tau)) : d^T \in d^{T*}, \tau \in (T, T]\}. \quad (9)$$

Using again the chain rule, we can verify that the closed loop description is indeed composed of the strategy and system descriptions:

$$p(d^T) = \prod_{\tau=T+1}^T s(y(\tau) | d^{T-1}, u(\tau)) r(u(\tau) | d^{T-1}). \quad (10)$$

Summary 1. A system model S given by (1) which fulfils the Estimation conditions A_e, B_e —which describes the controlled system for any causal strategy without any restriction on the control rate—has to specify the set of c.p.d.f.'s (9).

A causal strategy R is specified by the set of c.p.d.f.'s (8).

5. MODELLING AND CONTROL PERIOD

The general system model derived in previous section describes fully the system behaviour for any causal strategy. Thus, it fulfils Estimation conditions in the estimation phase, i.e. for $\tau < 1$.

Modelling, the necessary prerequisite of any estimation, has to be adapted to the intended use of the model. For the control-period estimation, we have to restrict the general system description (9) to be in a class which respects the Control conditions in the control phase, i.e. for $\tau > 0$.

The evident difference of the conditions A_c , B_c and A_r , B_r is the main obstacle we shall try to overcome here.

Dynamic programming is the most efficient conceptual procedure for the optimization under the causality restriction (2). We shall exploit it for restricting the system models (9) to those which respect also the Control conditions.

As a rule, the control horizon growing to infinity is assumed. Consequently we can (without substantial loss of generality) take T , T to be integer multiples of all inspected control periods.

Let, for a specific control period n , n data pairs be grouped

$$\begin{aligned} Y'(k) &= [y((k-1)n+1), y((k-1)n \\ &\quad + 2), \dots, y(kn)], \\ U'(k) &= [u((k-1)n+1), u((k-1)n \\ &\quad + 2), \dots, u(kn)]. \end{aligned}$$

The collection of all data on the time interval of interest reads

$$D^{T/n} = \{D(k)\}_{k=1}^{T/n+1}, \quad D(k) = (Y(k), U(k)).$$

Then, for the control design, the loss function (4) is approximately expressed in terms of the grouped data

$$L(d^T) \approx L_n(D^{T/n}) \quad (11)$$

where L_n is a suitable loss function.

The optimal inputs are (by A_c) required to be constant within the control period of the length n , i.e.

$$U(k) = \bar{u}(k) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \bar{u}(k)^n \mathbf{1}$$

with some scalar $\bar{u}(k)$ forming the system input under the control period n .

Under general conditions, the optimal strategy is deterministic. The optimal inputs $\{\bar{u}(k)\}$ are minimizing arguments of the dynamic programming equations. The equations are an analogy of the backward recursion for the expected loss

function (7)

$$\begin{aligned} V(D^{k-1}) &= \min_{u(k)} E[V(D^k) | D^{k-1}, \bar{u}(k)], \\ k &= \frac{T}{n}, \frac{T}{n} - 1, \dots, 1, \\ V(D^{T/n}) &= L_n(D^{T/n}). \end{aligned} \quad (12)$$

Formula (12) exemplifies the need for a system model yielding the c.p.d.f.

$$s(Y(k) | D^{k-1}, \bar{u}(k)) = s(Y(k) | D^{k-1}, U(k)). \quad (13)$$

Generally, the minimizing argument $U(k) = \bar{u}(k)^n \mathbf{1}$ is a function of the data Y^{k-1} , $U^{k-1} = \bar{u}^{k-1} \mathbf{1}$. In order to respect the condition (B_r), the vector sequence Y^{k-1} has to reduce to a scalar sequence \bar{y}^{k-1} . In terms of the c.p.d.f. (13), we have to search for a system model for which—under the control period n —holds

$$\begin{aligned} s(Y(k) | D^{k-1}, U(k)) &= s(Y(k) | Y^{k-1}, \bar{u}^k) \\ &= s(Y(k) | \bar{y}^{k-1}, \bar{u}^k) \\ &= s(Y(k) | \bar{d}^{k-1}, \bar{u}(k)) \end{aligned} \quad (14)$$

with $\bar{d}(k) = (\bar{y}(k), \bar{u}(k))$.

The restriction of the allowable system models (14) admits $\bar{y}(k)$ to be any function of data D^k . In the feedback, however, $\bar{y}(k)$ plays the rôle of a system output: a scalar-valued function of the output n -tuple measured within the k th control period. Thus, it is reasonable to take $\bar{y}(k)$ as a simple deterministic function $f(\cdot)$ of $Y(k)$ only

$$\bar{y}(k) = f(Y(k)). \quad (15)$$

Assuming that (14), (15) hold, we find that over the entire design horizon

$$V(D^k) = V(\bar{d}^k) \quad k = 1, 2, \dots, T/n$$

and that the optimal strategy fulfils the Control conditions.

Summary 2. The system model $\{s(y(\tau) | d^{T-1}, u(\tau))\}_{\tau=T/n+1}^T$ for which the equalities (14), (15) hold under any control strategy with control period n fulfils both the Estimation and the Control conditions.

6. SYSTEM MODEL ESTIMATION

The key feature of self-tuners is their ability to update the information about the system model throughout the control phase. Bayesian system identification supplies just the system model needed for the optimal control design sketched above. This fact is a decisive advantage which led us to adopt the Bayesian approach.

Let us briefly recall the Bayesian background needed in the sequel. The paper (Peterka, 1981) is a good reference for a detailed study.

6.1. Bayesian identification

In the Bayesian set-up, the c.p.d.f. $p(d^T)$ is assumed to be the marginal p.d.f. of the p.d.f. $p(d^T, \Theta)$ with an unknown (finite-dimensional) parameter $\Theta \in \Theta^*$.

Using the chain rule for c.p.d.f.'s and the relation of joint and marginal p.d.f.'s, we find that the needed p.d.f. can be expressed as

$$p(d^T) = \int \prod_{\tau=T+1}^T r(u(\tau) | d^{\tau-1}, \Theta) \times s(y(\tau) | u(\tau), d^{\tau-1}, \Theta) p(\Theta) d\Theta.$$

The c.p.d.f.'s $\{s(y(\tau) | u(\tau), d^{\tau-1}, \Theta)\}_{\tau=T+1}^T$ form the *parameterized* system model which describes the system response at time τ to the observed past process history if the parameter takes the value Θ .

The term "unknown" parameter means that Θ characterizes the properties of the system model unknown to the controller. Thus, the probability of possible values of $u(\tau)$ conditioned on the observed data $d^{\tau-1}$ and a parameter value Θ should not be influenced by Θ , i.e.

$$r(u(\tau) | d^{\tau-1}, \Theta) = r(u(\tau) | d^{\tau-1}), \quad \tau \in (T, T]. \quad (16)$$

Under these so called natural conditions of control (Peterka, 1981) the data observed up to time τ modify the expert's prior belief quantified by the p.d.f. $p(\Theta)$ to the p.d.f. $p(\Theta | d')$ according to the following form of the Bayes rule

$$p(\Theta | d') = \frac{l(\Theta | d') p(\Theta)}{l(\Theta | d') p(\Theta) d\Theta} \quad (17)$$

where the likelihood function $l(\Theta | d')$ is the product of the parametrized system models

$$l(\Theta | d') = \prod_{\tau=T+1}^T s(y(\tau) | u(\tau), d^{\tau-1}, \Theta) \quad (18)$$

for any $t \in (T, T]$.

Using again the chain rule, the relation between joint and marginal p.d.f.'s as well as the natural conditions of control (16), we find that the required system description (9) (Bayesian prediction) reads

$$\begin{aligned} & \{s(y(\tau) | u(\tau), d^{\tau-1}) \\ & = \int s(y(\tau) | u(\tau), d^{\tau-1}, \Theta) p(\Theta | d^{\tau-1}) d\Theta \end{aligned}$$

Note that no point estimation of unknown parameter is needed for determining the required system model (9).

6.2. Structure determination

The problem of structure determination arises when several $(1 < N < \infty)$ *structurally differing descriptions* $\{s(y(\tau) | u(\tau), d^{\tau-1}, \Theta)\}_{i=1}^N$ of a *single controlled system* are to be compared on a time interval. Individual members of the i th structure are generally distinguished by an unknown parameter $\Theta \in \Theta^*$.

Let $n \in \{1, 2, \dots, N\}$ denote the pointer to the appropriate (best, true ...) model structure

and let the compound parameter Θ be defined by

$$\Theta = \{n, \Theta^1, \dots, \Theta^N\} \in (\{1, \dots, N\}, \Theta^1, \dots, \Theta^N).$$

Then the following parameterized system model reflects the situation when we are uncertain about both the appropriate model structure n and the parameter Θ within it:

$$\begin{aligned} & s(y(\tau) | u(\tau), d^{\tau-1}, \Theta) \\ & = s(y(\tau) | u(\tau), d^{\tau-1}, n, \Theta^1, \dots, \Theta^N) \\ & = \prod_{i=1}^N s^{\delta_{in}}(y(\tau) | u(\tau), d^{\tau-1}, \Theta^i) \\ & \quad \begin{cases} s(y(\tau) | u(\tau), d^{\tau-1}, \Theta^1) & \text{if } n = 1, \\ s(y(\tau) | u(\tau), d^{\tau-1}, \Theta^2) & \text{if } n = 2, \\ \dots \\ s(y(\tau) | u(\tau), d^{\tau-1}, \Theta^N) & \text{if } n = N. \end{cases} \end{aligned}$$

In the product form of the above formula, the letter δ_{in} denotes the Kronecker symbol.

Note that this way of modelling is very flexible due to the possibility of joining models of a quite different nature.

Assuming prior belief to be assigned independently to particular entries of Θ , the Bayes rule specializes to

$$\begin{aligned} & p(n, \Theta^1, \dots, \Theta^N | d') \\ & = \frac{\prod_{i=1}^N l^{\delta_{in}}(\Theta^i | d') p(\Theta^i)}{\sum_{j=1}^N \int l^{\delta_{jn}}(\Theta^j | d') p(\Theta^j) d\Theta^j p(j)} p(n) \end{aligned}$$

where the likelihood $l(\Theta | d')$ assigned to an i th model structure is defined by (18) written for the i th parametrized system model.

By inspecting this formula, it can be seen that the parameter Θ within the i th structure is identified irrespectively of the other structures assumed.

Formally, no estimate of the proper structure is needed when constructing the system model. This is clear from the imbedding of the structure estimation into the parameter estimation framework and from the conclusion of the previous Section. We are, however, often forced (usually because of computational reasons) to select a single structure at the end of the estimation phase. Then the parametrized model of the selected structure is identified within the control phase.

In spite of the diversity of the parametrization used, the identified submodels are quite comparable through the marginal posterior probability $p(n | d')$.

Summary 3. All information needed for selecting a single structure at the end of the estimation phase ($\tau=0$) is contained in the marginal

posterior probability

$$p(n | d^n) = \frac{\int l(\theta | d^n) p(\theta) d^n \theta}{\sum_{i=1}^n \int l(\theta | d^n) p(\theta) d^n \theta p(i)} p(n). \quad (19)$$

7 CONTROL-RATE DEPENDENT LINEAR NORMAL REGRESSION MODEL

Among linear input-output models, the ARMAX model is the most general description of a linear stochastic system. The majority of applied self-tuners rely, however, on a regression model (often working on prefiltered data). Popularity of the regression models stems mainly from the fact that their identification leads to recursively computable least squares. For the same reason, we search for control-rate dependent system descriptions within the class of regression models.

Control strategies used in self-tuning are often based on an enforced separation of the identification and control synthesis (certainty equivalence). We restrict ourselves to this case, too.

We shall interpret the searched control period n as the model structure. Then, according to Summary 3, the posterior probability $p(n | d^n)$ compresses all information needed for a prior choice of the control rate. In order to evaluate it, control-rate dependent parametrized system models have to be specified.

The resulting system models (Bayesian predictions) should fulfil the conditions (14), (15) in Summary 2. For certainty-equivalent controllers, it is sufficient to ask validity of (14), (15) for the parametrized system models only.

We claim that the following regression model fulfils the discussed conditions and it is also technically sound. The output n tuples are modelled by

$$Y(k) = {}^n\mathbf{1} \left[\sum_{i=1}^{n_l} {}^na_i \bar{y}(k-i) + \sum_{i=1}^{n_b} {}^nb'_i U(k-i) + {}^nc \right] + E(k). \quad (20)$$

The function (15) is defined by the formula

$$\bar{y}(k) = \frac{1}{n} \sum_{i=1}^n Y_i(k) = \frac{1}{n} \sum_{i=1}^n y((k-1)n+i). \quad (21)$$

The meaning of the other quantities is: nl_y , nl_u are respective orders; na_i are scalar autoregression coefficients; nb_i are vector regression coefficients; nc is a common absolute term; and $\{E(k)\}$ denotes white normal noise, i.e. the sequence of independent normally distributed stochastic variables with the zero expectation and the common precision matrix

(inversion of the covariance matrix)

$${}^n\Omega_E = {}^n\Omega \left[{}^n\omega \frac{{}^n\mathbf{1}\mathbf{1}'}{n} + I - \frac{{}^n\mathbf{1}\mathbf{1}'}{n} \right]. \quad (22)$$

In Formula (22), I denotes unit matrix, ${}^n\Omega$ and ${}^n\omega$ are positive scalars.

Normality and whiteness of the noise imply that the required parametrized system model of the controlled system is specified. It reads

$$s(Y(k) | Y^{k-1}, U^k, {}^n\Theta) = \mathcal{N}_{Y(k)}({}^n\mathbf{1}z'(k){}^nP, ({}^n\Omega_E)^{-1})$$

where the normal c.p.d.f. \mathcal{N}_Y of Y having an expectation M and a precision matrix Ω_E takes the familiar form

$$\mathcal{N}_Y(M, \Omega_E^{-1}) = \frac{|\Omega_E|^{1/2}}{2\pi} \exp \left[-\frac{1}{2} (Y - M)' \Omega_E (Y - M) \right]. \quad (23)$$

The assumed precision matrix is defined by (22). The expectation—defined by Equation (20)—can be written in the compact form

$$M = {}^n\mathbf{1}z'(k){}^nP$$

with

$${}^nz'(k) = [U'(k), \dots, U'(k - {}^nl_u), \bar{y}(k-1), \dots, \bar{y}(k - {}^nl_y), 1],$$

$${}^nP' = [{}^nb'_0, {}^nb'_1, \dots, {}^nb'_{l_u}, {}^na_1, {}^na_2, \dots, {}^na_{l_y}, {}^nc].$$

The vectors nz , nP have the dimension

$${}^nm = ({}^nl_u + 1)n + {}^nl_y + 1 \quad (24)$$

The unknown model parameter is ${}^n\Theta = [{}^nP, {}^n\Omega, {}^n\omega]$.

Using a simple algebra, the introduced parameterized system model can be rewritten into the form

$$s(Y(k) | Y^{k-1}, U^k, {}^nP, {}^n\Omega, {}^n\omega) = \left(\frac{{}^n\Omega}{2\pi} \right)^{n/2} ({}^n\omega)^{1/2} \times \exp \left\{ -\frac{{}^n\Omega}{2} \left[\frac{Y'(k)Y(k)}{n} - \bar{y}^2(k) + {}^n\omega \left[\begin{matrix} -1 \\ {}^nP \end{matrix} \right]' \left[\begin{matrix} \bar{y}(k) \\ {}^nz(k) \end{matrix} \right] \left[\begin{matrix} \bar{y}(k) \\ {}^nz(k) \end{matrix} \right]' \left[\begin{matrix} -1 \\ {}^nP \end{matrix} \right] \right] \right\} \quad (25)$$

which will help us in describing its identification.

The model adds no restriction on inputs and together with the strategy description specifies the joint p.d.f. of all data measured within the estimation phase. Thus, it meets the Estimation conditions.

The direct inspection of the model (25) verifies the identity

$$s(Y(k) | Y^{k-1}, U^k, {}^n\Theta) = s(Y(k) | \bar{y}^{k-1}, U^k, {}^n\Theta).$$

If the input changes with the rate n then its influence on the output reduces to

$${}^n\mathbf{1} \sum_{i=0}^{n_b} {}^nb_i \bar{u}(k-i)$$

where the scalar coefficients nb_i are defined

" $\bar{b}_i = "b_i"$ 1. Thus, in the control phase, the Control conditions $s(Y(k) | \bar{y}^{k-1}, U^k, " \Theta) = s(Y(k) | \bar{d}^{k-1}, \bar{u}(k), " \Theta)$ are met.

The following remarks should confirm that the model (25) is technically sound.

The mean value $\bar{y}(k)$ (21) minimizes the quadratic form

$$\sum_{i=1}^n (Y_i(k) - \bar{y}(k))^2 = \sum_{i=1}^n (y((k-1)n + i) - \bar{y}(k))^2.$$

For the quadratic control design, this simple property makes $\bar{y}(k)$ preferable function of the output n -tuple in k th control period: it produces a reasonable approximation of the quadratic loss L by the quadratic loss L_n adapted to the control period n (11).

The mean has been invented by adopting the idea of piece-wise filtering (Kárný *et al.*, 1988) which can be exploited further.

The degenerate form of the coefficients at the inputs respects the basic need for converting the model into a single-input one in the control phase. At the same time, the coefficients weight the individual inputs in the estimation phase.

The reduction (14) is achieved by assuming the common contribution of \bar{y}^{k-1} and U^k to the prediction of $Y(k)$ -entries. Intersample behaviour is modelled by a strong correlation of noise entries.

The chosen precision matrix has $n-1$ unit eigenvalues and one equal to " ω " which corresponds to the eigenvector " $\mathbf{1}$ ". If $1 \gg \omega$ then there is a ridge on the parametrized system model (25) assigning high probabilities to $Y(k) = \mu \mathbf{1}$ (μ a real scalar). The closer μ is to " z/P " the higher the probability is. This property is clearly necessary for a successful control with the control period n .

Summary 4. For a given parameter " $\Theta = (P, \Omega, \omega)$ ", the parametrized system model (25) meets both the Estimation and the Control conditions.

8. IDENTIFICATION OF CONTROL RATE DEPENDENT REGRESSION MODEL

The identification of the fixed-structure model is an intermediate step for structure determination. We shall perform it for the regression models introduced in Section 7.

Using the system model (25), the likelihood function (18) at a time instant $\tau = k(\tau)n - k(\tau)$ an integer number—reads (with superscript n suppressed)

$$l(\Theta | D^{k(\tau)}) = \left(\frac{\Omega}{2\pi} \right)^{v(\tau)/2} \omega^{k(\tau)/2} \times \exp \left\{ -\frac{n\Omega}{2} \left[\lambda(k(\tau)) + \omega \begin{bmatrix} -1 \\ P \end{bmatrix}' V(k(\tau)) \begin{bmatrix} -1 \\ P \end{bmatrix} \right] \right\} \quad (26)$$

where (with τ suppressed)

$$V(k) = V(k-1) + \begin{bmatrix} \bar{y}(k) \\ z(k) \end{bmatrix} \begin{bmatrix} \bar{y}(k) \\ z(k) \end{bmatrix}' = \begin{bmatrix} V_{yy}(k) & V_{zy}(k) \\ V_{zy}(k) & V_{zz}(k) \end{bmatrix}. \quad (27)$$

In the block form of $V(\cdot)$, $V_i(\cdot)$ is a scalar. The scalar $\lambda(\cdot)$ is defined by

$$\lambda(k) = \lambda(k-1) + \frac{Y'(k)Y(k)}{n} - \bar{y}^2(k). \quad (28)$$

The scalar $v(\tau)$, defining $k(\tau)$ with the help of integer division \div , counts the number of data items

$$v(\tau) = v(\tau-1) + 1, \quad k(\tau) = v(\tau) \div n. \quad (29)$$

All recursions start from zero initial conditions.

Choosing a self-reproducing prior p.d.f. (De Groot, 1970), the constructed posterior p.d.f. exhibits the same functional form as the likelihood function. Prior belief of the user is expressed by nonzero initial conditions $V(T/n)$, $\lambda(T/n)$, $v(T/n)$ of the above recursions.

Normalization of the function obtained to the p.d.f. has to be possible. It is guaranteed if at the initial time instant T (cf. (24))

$$V(T/n) > 0, \quad \lambda(T/n) > 0, \quad v(T/n) > n(m-2). \quad (30)$$

The matrix inequality $V > 0$ means that V is positive definite.

Proposition 1. For a given control period n , time τ which is its integer multiple, the initial conditions (30) and under natural conditions of control (16), the Bayesian estimate of the parameter $\Theta = (P, \Omega, \omega)$ specifying the regression model (25) takes the form

$$p(P, \Omega, \omega | d^1, n) = \frac{l(\Theta | d^1)}{\mathcal{J}[V(v(\tau) \div n), \lambda(v(\tau) \div n), v(\tau), n)]}. \quad (31)$$

The likelihood function $l(\cdot | \cdot)$ (modified by the prior p.d.f.) is given by Formula (26). The normalizing factor reads

$$\mathcal{J}[V, \lambda, v, n] = \psi(v, n) \lambda^a \bar{\lambda}^{m/2-a} |V_z|^{-1/2}. \quad (32)$$

The quantities V, λ, v are sufficient statistics for estimating P, Ω, ω . They are specified by Equations (27), (28) and (29) and related to $\bar{\lambda}, a, b, \psi$ as follows

$$\lambda = V_y - V_{zy}' V_z^{-1} V_{zy}, \quad a = \frac{v}{2n} + 1, \quad b = \frac{v}{2} + 1, \quad (33)$$

$$\psi(v, n) = \pi^{1-b} H_m \Gamma \left[\frac{m}{2} \right] n^{-b} \Gamma \left[a - \frac{m}{2} \right] \Gamma[b-a],$$

with Γ denoting Euler function and H_m volume of the unit ball in m -dimensional space.

The formulae are formally valid for $n=1$ if

the factors depending on $b - a = 0$ are set equal to 1.

Proof. The proof adds to the formerly derived results just the evaluation of the integral of the likelihood function (see Appendix).

9. MAIN RESULT: CONTROL-RATE ESTIMATION

With the preparation made, the Bayesian estimate of the control period is obtained by specializing the general structure-estimation formula (19) to the regression model assumed.

Proposition 2. For T being an integer multiple of all competitive control periods $n \in \{1, 2, \dots, N\}$ and under natural conditions of control (16), the Bayesian estimate of the structure n pointing to the regression models (25) takes the form

$$p(n \mid d^0)$$
$$= \kappa p(n) \cdot \frac{\mathcal{J}[{}^nV(v \div n), {}^n\lambda(v \div n), v, n]}{\mathcal{J}[{}^nV, {}^n\lambda, v, n]}$$
$$= \kappa p(n) \frac{\psi(v, n)}{\psi(v, n)} \frac{{}^n\lambda^{a-b} \cdot {}^n\tilde{\lambda}^{m/2-a}}{{}^n\lambda^{a-b} \cdot {}^n\tilde{\lambda}^{m/2-a}} \left(\frac{|{}^nV_z|}{|{}^nV_z|} \right)^{-1/2}$$

(i)

(ii)

(iii)

(iv)

$$(34)$$

where

the quantities $V, \lambda, \tilde{\lambda}, v, a, b$ have been defined in connection with Proposition 1, however, their dependence on n is made explicit here;

the statistics $V_z, \lambda, \tilde{\lambda}, a, b$ are related to the end of the estimation phase and their boldfaced versions to its beginning;

κ is a normalizing factor independent of n ; item $p(n)$ is the p.d.f. (with respect to counting measure) describing prior user's belief;

$$\psi(v, n) = n^{-b} \Gamma[b - a] \Gamma\left[{}^na(0) - \frac{{}^nm}{2}\right].$$

Proof. The proof is obtained by a straightforward combination of Equations (17), (19), (31), (32) and (33), noticing that integral in the numerator of the ratio (19) can be written as the ratio of the integrals \mathcal{J} evaluated at the end and the beginning of the estimation phase, respectively. □

The labelled factors in (34) correspond to the basic ingredients which modify the prior belief to the inspected control periods. Specifically, it is influenced by

- (i)

(ii)

(iii)
- the number of data available v , the "aggregation" degree n and the number of estimated parameters nm (24);

the intersampling dispersion reflected in λ (the exponent is an increasing function of n);

the ability to predict the output mean (\bar{y}) reflected in the corresponding remainder of

least squares $\tilde{\lambda}$ (the exponent is proportional to the number of data: it has the strongest discarding power);

- (iv)
- the uncertainty of the estimated parameters (the exponent is fixed: this "second-order" quantity decides among models with roughly same predictive potential).

10. IMPLEMENTATION ASPECTS

Choice of the prior p.d.f.'s $p(n), p({}^n\Theta)$. Usually, not too much information is available when the control period is selected. Thus, it makes sense (at time T):

—to select a uniform distribution over the compared periods

$$p(n) = \frac{1}{N};$$

—to choose $V_{zy} = 0, V_z = \varepsilon I, 0 < \varepsilon \ll 1$ which corresponds to a very flat distribution of the regression coefficients P ;

—to respect the relation $0 < \lambda \ll \tilde{\lambda} = V_z$ which is necessary for $\mathcal{E}[\omega] \ll 1$ (see Section 7);

—to set $v = n(m - 2) + \varepsilon$ which guarantees a proper but flat distribution of Ω .

A specific prior knowledge can be introduced in the way outlined in (Kárný, 1984).

Evaluation of $p(n \mid d^0)$, (34). When programming this formula we have found it useful: to use its logarithmic version in order to prevent overflows; to apply the Stirling formula for approximation of the Γ -function; to normalize rough data to a zero mean and a common dispersion—numerical problems are avoided when compressing the data into the matrix V ; to exploit the regressor shifting structure when collecting the data into the matrix V in order to spare the computation time (Kárný, 1983); and to use LD (or Choleski) factorization (Bierman, 1977) of V for the efficient evaluating $|V_z|$ and $\tilde{\lambda}$.

Concerning the last item the key observations are: if $V = F'F$ where F is a lower triangular matrix with positive diagonal terms F_1, F_2, \dots, F_{m+1} then

$$\tilde{\lambda}^{1/2} = F_1 \quad \text{and} \quad |V_z|^{1/2} = \prod_{i=2}^{m+1} F_i.$$

11. ILLUSTRATIVE EXAMPLE

The proposed algorithm has been implemented and tested within a flexible experimental environment SIC [package for Simulation, Identification and adaptive Control of stochastic systems, (Kárný *et al.*, 1985; Kulhavý, 1988)]. The LQG self-tuner used for cross-validation of the gained results is described in (Kárný *et al.*, 1985). For illustration, we present results of a single test.

TABLE 1. LOG-LIKELIHOODS ASSIGNED TO THE INSPECTED CONTROL PERIODS (BOX MARKS MAXIMUM LIKELIHOOD ESTIMATE)

n	1	2	3	4	5	6
$\ln(l(n d_{119}^0))$	-189	-186	-210	-286	-369	-493
$\ln(l(n d_{500}^0))$	-891	-662	-571	-571	-685	-766

Estimation phase

Simulated system.

$$\begin{bmatrix} y(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} 0.98 & 1 \\ 0 & -0.9 \end{bmatrix} \begin{bmatrix} y(t-1) \\ x(t-1) \end{bmatrix} + \begin{bmatrix} 0.01 \\ 0.11 \end{bmatrix} u(t) + \begin{bmatrix} 0.01 \\ -0.0095 \end{bmatrix} u(t-1) + \begin{bmatrix} \zeta_1(t) \\ \zeta_2(t) \end{bmatrix}$$

where

$x(t)$ is an unobserved inner variable, completely neglected both in the estimation and control phases;

$\zeta_1(t)$, $\zeta_2(t)$ are mutually independent zero mean normal random variables with the common dispersion 0.1.

Input signal. Zero mean white noise with dispersion 0.1 and independent of ζ .†

Estimated structure, prior data and tested periods. First order model $l_v = l_u = 1$ is assumed, i.e. the model order is underestimated.

The control periods $n \in \{1, 2, \dots, 6\}$ are tested. The recommended initial values of statistics are used.

Results of the estimation phase

The posterior probabilities of the control periods are: $p(1 | d_{199}^0) = 0.0474$, $p(2 | d_{199}^0) = 0.9526$ and zero otherwise; and $p(3 | d_{500}^0) = 1.0000$ and zero otherwise.

The detailed results can be seen from the Table 1 where logarithm of the likelihoods $l(n | d_{T+1}^0) = \text{const.}$ $p(n | d_{T+1}^0)$ are listed.

The maximum *a posteriori* likelihood estimates (marked by boxes) are clearly optimal for a majority of loss functions which could be selected for a point estimate of n .

Control phase and cross-validation

Optimized criterion.

$$\mathcal{K} = \lim_{T \rightarrow \infty} \mathbb{E}[L(d^T)],$$

$$L(d^T) = \frac{1}{T} \left[\sum_{t=1}^T y(t)^2 + u(T)^2 + u(T-1)^2 \right]$$

System, system model and tested periods. The same system model and the first order regression model are used in simulating the control phase. The control periods $n = 1, 2, \dots, 6$ are tested.

TABLE 2. LOSSES ACHIEVED FOR THE INSPECTED CONTROL PERIODS (BOX MARKS MINIMUM LOSS, ∞ INDICATES UNSTABILITY)

n	1	2	3	4	5	6
$L(d_{121}^{120})$	∞	0.815	1.086	9.619	5.260	4.036
$L(d_{600}^{600})$	∞	0.638	0.648	2.520	1.747	1.504
$L(d_{121}^{600})$	∞	0.592	0.538	0.745	0.868	0.871

Prior information, control strategy. Rather vague prior knowledge is used by the controller (zero prior mean of the unknown parameters, and a large diagonal covariance matrix). Certainty equivalence version of IST strategy (Kárný *et al.*, 1985) with 20 iteration steps per the control period is applied.

Results of control phase

The achieved sample values of the loss function are summarized in the Table 2, where ∞ indicates unstable closed loop and $L(d_{121}^{600})$ denotes sample loss in the interval [121,600] (start-up behaviour omitted). The achieved values illustrate the strength of the proposed estimation algorithm.

12. CONCLUSIONS

The paper presents an attempt to find well-founded algorithm for a data-based choice of the control period. The theory of Bayesian structure determination proves to be natural candidate for solving this task. In addition to the widely known advantages (Peterka, 1981), the Bayesian set-up can be connected with approximation theory (Harris and Heindel, 1979). This fact is of special importance in the solved problem because its non-trivial solution exists under misspecification only.

Within the Bayesian theory, the proposed algorithm simply compares posterior probabilities of alternative system descriptions. When searching for such alternative system descriptions, simple but important circumstances are faced: learning data are collected and inputs are changed with a rate which differs from the one searched for; and the found model structure has to reduce to SISO system description when designing the controller with a fixed control rate.

A subclass of regression models has been proposed which respects these circumstances. The subclass is by no means unique. The chosen version has been influenced by intention not to change the existing design of LQG self-tuners to which the discussed choice should serve.

Up to now, a couple of various simulation cases have been tested with the following common features observed:

- (1) "dangerous" control rates have been recognized as a rule;

† Recall the convention made: the necessary step of the transport delay is included into the controller.

- (2) maximum of the posterior p.d.f. $p(n | d^0)$ is attained nearby the reasonable control rate if such a control period exists; and
- (3) an overestimation of n has occurred for long learning data sequences (≈ 5000) [forgetting in the manner of (Kárný and Kulhavý, 1988) should be used].

In summary, the results are encouraging in spite of evident incompleteness of the open-loop based comparison. The prediction of closed loop behaviour in the vein of (Kárný *et al.*, 1990) seems to be proper way for improvement of the proposed method.

REFERENCES

- Bierman, G. J. (1977). *Factorization Methods for Discrete Sequential Estimation*. Academic Press, New York.
- De Groot, M. H. (1970). *Optimal Statistical Decision*. McGraw-Hill, New York.
- Harris, B. and G. Heinel (1979). The relation between statistical decision theory and approximation theory. In Rustagi, J. G. (Ed.), *Optimizing Methods in Statistics*. Academic Press, New York, 263–272.
- Kárný, M. (1983). Algorithms for determining the model structure of a controlled system. *Kybernetika* **19**, 164–178.
- Kárný, M. (1984). Quantification of prior knowledge about global characteristics of linear normal model. *Kybernetika* **20**, 376–385.
- Kárný, M., A. Halousková, J. Böhm, R. Kulhavý and P. Nedoma (1985). Design of linear quadratic adaptive control: theory and algorithms for practice. *Kybernetika* **21** (Suppl. to numbers 3, 4, 5, 6).
- Kárný, M. and R. Kulhavý (1988). Structure determination of regression-type models for adaptive prediction and control. In Spall, J. C. (Ed.), *Bayesian Analysis of Time Series and Dynamic Models*. Marcel Dekker, New York.
- Kárný, M., A. Halousková and I. Nagy (1988). Modelling, identification and adaptive control of cross-direction basis weight of paper sheets. *Int. Conf. Control* **88**, Oxford, 159–164.
- Kárný, M., T. Jeníček and W. Ottenheimer (1990). Contribution to prior tuning of LQG self-tuners. *Kybernetika* **26**, 107–121.
- Kulhavý, R. (1988). *SIC: User's Guide* (version 1.1). Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, Prague.
- Peterka, V. and M. Kárný (1979). Bayesian system classification. *Preprints of 5th IFAC Symp. on Identification and System Parameter Estimation*. Darmstadt, **1**, 349–356.
- Peterka, V. (1981). Bayesian system identification. In: Eykhoff, P. (Ed.), *Trends and Progress in System Identification*. Pergamon Press, Oxford, 239–304.
- Rao, M. M. (1987). *Measure Theory and Integration. Pure and Applied Mathematics*. Wiley-Interscience, New York.

APPENDIX: INTEGRATION IN PROPOSITION 1

For completeness, the evaluation of the integral $\mathcal{J}(V, \lambda, v, n)$ needed in Proposition 1 is sketched. To spare the space, the substitutions and definitions of auxiliary quantities are written within the string of equalities in norm-like brackets $\|\cdot\|$, new variable is called x and a single integral sign is used (ω, Ω are non-negative scalars and P is m -dimensional real vector).

Except a sequence of substitutions, the completion of squares of a quadratic form in P is used with the block decomposition of the positive definite matrix

$$= \begin{bmatrix} V_y & V_{yz} \\ V_{zy} & V_z \end{bmatrix}.$$

Moreover, the following definitions of the Euler functions are used

$$\Gamma[a] = \int_0^\infty x^{a-1} \exp(-x) dx$$

$$B[a, b] = \int_0^1 x^{a-1} (1-x)^{b-1} dx = \frac{\Gamma[a]\Gamma[b]}{\Gamma[a+b]}.$$

Evaluation of I

$$\mathcal{J}[V, \lambda, v, n] =$$

$$\begin{aligned} & \int \left(\frac{\Omega}{2\pi} \right)^{v/2} \omega^{v/2} \exp \left\{ -\frac{n\Omega}{2} \left[\lambda + \omega \begin{bmatrix} -1 \\ P \end{bmatrix}' V \begin{bmatrix} -1 \\ P \end{bmatrix} \right] \right\} dP d\Omega d\omega \\ &= \left\| \gamma \approx \begin{bmatrix} -1 \\ P \end{bmatrix}' V \begin{bmatrix} -1 \\ P \end{bmatrix}; a \approx \frac{v}{2n} + 1; b \approx \frac{v}{2} + 1; x = \frac{\Omega n \gamma}{2} \right\| \\ &= (2\pi)^{-1} \binom{n}{2} \int \Gamma[a] \gamma^{-a} \Omega^{b-a-1} \exp \left[-\frac{\Omega \lambda n}{2} \right] dP d\Omega \\ &= \left\| x = \frac{\lambda n}{2} \Omega \right\| = (2\pi)^{-1} \binom{n}{2} \int \Gamma[a] \Gamma[b-a] \lambda^a \gamma^{-b} \int \gamma^{-a} dP \\ &= \left\| \gamma \approx \tilde{\lambda} + (P - \tilde{P})' V_z (P - \tilde{P}); \tilde{\lambda} \approx V_y - V_{zy} V_z^{-1} V_{yz}; \right. \\ & \quad \left. P \approx V_z^{-1} V_{yz}; x = \lambda^{-1/2} V_z^{1/2} (P - \tilde{P}) \right\| \\ &= (2\pi)^{-1} \binom{n}{2} \int \Gamma[a] \Gamma[b-a] \lambda^a \gamma^{-b} \tilde{\lambda}^{-a+m/2} |V_z|^{-1/2} \\ & \quad \times \int (1+x^2)^{-a} dx \\ &= \|H_m\| \approx \text{volume of the } m\text{-dimensional unit ball}; \\ & \quad x \text{ is transformed to polar coordinates with radius } r\| \\ &= (2\pi)^{-1} \binom{n}{2} \int \Gamma[a] \Gamma[b-a] \lambda^a \gamma^{-b} \tilde{\lambda}^{-a+m/2} |V_z|^{-1/2} H_m \\ & \quad \times \int_0^1 r^{m-1} (1+r^2)^{-a} dr \\ &= (2\pi)^{-1} \binom{n}{2} \int \Gamma[a] \Gamma[b-a] \lambda^a \gamma^{-b} \tilde{\lambda}^{-a+m/2} |V_z|^{-1/2} \\ & \quad \times H_m \frac{1}{2} B \left[\frac{m}{2}, a - \frac{m}{2} \right] \\ &= \pi^{-1} \binom{n}{2} H_m \Gamma \left[\frac{m}{2} \right] n^{-b} \Gamma \left[a - \frac{m}{2} \right] \Gamma[b-a] \lambda^a \gamma^{-b} \tilde{\lambda}^{-a+m/2} |V_z|^{-1/2}. \end{aligned}$$

□

A Geometric Approach to Proportional-plus-derivative Feedback Using Quotient and Partitioned Subspaces*

V. L. SYRMOS† and F. L. LEWIS‡

A geometric theory to proportional-plus-derivative feedback using quotient and partitioned subspaces, as well as spectrum assignability techniques using nonsquare pencils, provides advantageous computational methods for singular as well as state-variable systems.

Key Words—Singular systems, proportional-plus-derivative feedback, geometric theory.

Abstract—In this paper a new characterization of invariant subspaces is presented, using the notions of partitioned and quotient subspaces. This classification is based on a feedback approach that exhibits the importance of these subspaces for the problem of proportional-plus-derivative feedback. It also provides the ability to decompose the closed-loop system into two subsystems that completely characterize the closed-loop behavior. A feedback-free formulation is also considered which opens new horizons for the problem of general semistate feedback by utilizing nonsquare pencils for spectrum assignability. The importance of this technique arises from the fact that we deal with reduced order pencils, therefore the computational methods are more stable. In order to accomplish this, we use two generalized Lyapunov equations and exploit the generalized Hessenberg form.

1. INTRODUCTION

THE CLASSIFICATION and characterization of invariant subspaces has always been an important topic in control theory. In state-variable systems (Basile and Marro, 1969; Willems, 1981, 1982; Wonham, 1979) these ideas constitute a powerful tool for the study of pure proportional feedback (Wonham, 1979). Later, matrix pencil characterizations for invariant subspaces were presented in the literature (Jaffe and Karcnias, 1981). These characterizations bring together the matrix pencil (Gantmacher, 1959) and geometric theories for state-variable systems.

For singular systems (Lewis, 1986; Özçaldıran, 1985) various approaches were used for the characterizations of invariant subspaces (Lewis, 1986; Lewis and Özçaldıran, 1989; Lewis and Symos, 1989; Malabre, 1987; Özçaldıran, 1985, 1986). Moreover, an attractive theory for matrix pencil theory was presented (Karcnias and Kalogeropoulos, 1987; Van Dooren, 1979, 1981). These characterizations led to a complete theory for the use of pure proportional feedback. In Shayman and Zhou (1987) a classification of singular systems using constant ratio proportional-plus-derivative feedback was presented. This approach, although it is an attractive one, is based on nongeneral proportional-plus-derivative feedbacks and does not present general geometric characterizations for the involved invariant subspaces.

This paper aims to provide a complete feedback characterization and classification of invariant subspaces for singular systems, based on the notions of partitioned and quotient subspaces. The definition of *partitioned* subspace that is presented in this paper is closely related to the notions of quotient subspaces. Partitioned subspaces are the primary tool for proportional-plus-derivative feedback.

In Section 2 we introduce some geometric notions that will be of great help. It is shown that a partitioned subspace defines a class of quotient subspaces. These quotient subspaces define a set of induced maps that carry the information of the original system and is called an *induced partition*. This class of subspaces satisfies certain properties with respect to the induced maps. The motivation for this approach comes from the matrix pencil theory, since these

* Received 23 August 1989 revised 1 February 1990, received in final form 9 April 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor P. Dorato under the direction of Editor H. Kwakernaak.

† School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

‡ Author to whom all correspondence should be addressed.

maps constitute the idea for the definition of the induced pencils. These pencils play a role of prime importance for the proportional-plus-derivative feedback characterization of these subspaces.

The concept of a *regular* partition is also presented in this section. This notion is substantial, since it is the one that guarantees the closed-loop regularity of the system. It is noteworthy that we do not assume the regularity property for the open-loop system.

Moreover, in this section an alternative approach, that is, a feedback-free description of the original system is presented. This formulation characterizes a new class of *feedback-free induced partitions*. These partitions also define a set of induced maps; moreover, these maps provide a feedback-free description of the system using the corresponding induced pencils. It is important to mention at this point that these pencils are nonsquare and invariant under the presence of B . The use of nonsquare matrices constitutes a new concept in control theory even in the state-variable case. This advantage is discussed later in the paper.

In Section 2 we study the closed-loop system restricted to a regular partitioned subspace. Our main interest here is to present a technique for pole placement using general proportional-plus-derivative feedback. It is noteworthy that state-variable systems are not closed under the action of proportional-plus-derivative feedback. On the other hand, singular systems do not share this drawback. Consequently, a complete geometric theory can be presented for these kinds of feedbacks by exploiting the theory of singular systems. This will be accomplished by utilizing the geometric notions in Section 2.

Using the results of the induced partitions we decompose the closed-loop system into two subsystems. Specifically, in one subsystem proportional feedback is used while in the other derivative feedback is used. During that decomposition our main concern focuses on the regularity of the closed-loop system and how it is preserved under this decomposition. Subsequently we present some ancillary results which guarantee that the closed-loop regularity for the original system is preserved under that decomposition for a class of constructed induced partitioned subspaces. This latter will be called a *well-defined* induced partition. In other words a well-defined induced partition does not destroy the closed-loop regularity of the system. Once we decompose the closed-loop system, we use two generalized Lyapunov equations to find the feedback that assigns the desired poles of the two subsystems while guaranteeing the closed-

loop regularity. The solution to this equation is based on the generalized Hessenberg form (Verhagen and Van Dooren, 1986), and exploits the use of unitary transformations. Finally we show that the gains that assign the poles for these two subsystems assign the same desired poles for the original closed-loop system.

Under these concepts we proceed in Section 4 to relate the spectrum assignability problem to feedback-free induced partitions. This section opens new horizons, not only in singular systems but also in state-variable systems for the confrontation of this problem using nonsquare matrices. Here we relate the geometric concepts of the feedback-free partitions and quotient subspaces.

Initially in Section 4 we present some ancillary results that will help us in our exploration. We show that any nonsquare pencil restricted to a partition can be decomposed into two feedback-free induced pencils. Moreover we can find an equivalence relation between the closed-loop system pencil and the feedback-free induced pencils. This result immediately focuses our interest on the use of nonsquare pencils for the spectrum assignability problem. In light of these concepts we restrict the feedback-free to a regular partition and show how to construct a *well-defined feedback-free* induced partition. Specifically Theorem 13 reveals how a well-defined partition is related to a well-defined feedback-free partition in terms of the decomposed subsystems. This result is of great importance since it constitutes the key for the regularity of the closed-loop system.

Having guaranteed the closed-loop regularity of the system, we next discuss assigning the poles using the nonsquare pencils. This is accomplished using the feedback-free generalized Lyapunov equations. We point out that these equations are of reduced order. The solution of these equations is also based on the generalized Hessenberg form that exploits all of the above advantages. Moreover, in the feedback-free formulation they do not depend on B . Finally, we show how to reconstruct our original closed-loop system and how to select the feedback gains that assign the desired modes and guarantee the closed-loop system regularity. For this procedure we provide an algorithm.

2. SOME GEOMETRIC NOTIONS FOR QUOTIENT SUBSPACES

In this section we present some geometric concepts for quotient subspaces. As we shall see in Section 3, these concepts will lead us to a complete theory for the use of proportional-plus-derivative feedback (PD) in singular systems.

Two closely related approaches to the geometric characterization of these subspaces that exploit the notions of quotient subspaces are considered. These two formulations are presented in Theorem 1 and Lemma 4. First it is necessary to give some ancillary results that will help us in our attempt.

Consider the generalized, or singular, linear dynamical system

$$E\dot{x} = Ax + Bu \quad (2.1)$$

where $x \in \mathbb{R}^n$, and $u \in \mathbb{R}^m$, and E square and generally singular.

We will call (2.1) *regular*, if

$$\Delta(s) = \det(sE - A) \neq 0. \quad (2.2)$$

Regularity is equivalent to the existence and uniqueness of the solution of $x(t)$ given $x(0^-)$ and $u(t)$. The roots of $\Delta(s)$ are called the *finite (relative) eigenvalues* of (E, A) . These are simply the finite zeros of the pencil $(sE - A)$. The *(relative) spectrum* of (E, A) is the union of the finite and infinite zeros of $(sE - A)$. We denote by $\sigma(E, A)$ the finite spectrum of (E, A) . The spectrum of a single matrix F we denote by $\sigma(F)$.

In this paper we do not assume the property regularity [i.e. (2.2)] which is too restrictive.

2.1. Quotient and partitioned subspaces

Define $\mathcal{F} \subset \mathbb{R}^n$ as an (A, E, \mathcal{B}) -invariant subspace for (2.1) if it satisfies

$$A\mathcal{F} \subset E\mathcal{F} + \mathcal{B}. \quad (2.3)$$

These subspaces have been used in connection with proportional feedback by Lewis and Özçaldıran (1989).

Similarly, define $\mathcal{F} \subset \mathbb{R}^n$ as an (E, A, \mathcal{B}) -invariant subspace for (2.1) if it satisfies

$$E\mathcal{F} \subset A\mathcal{F} + \mathcal{B}. \quad (2.4)$$

These subspaces have been used in connection with derivative feedback by Lewis and Syrmos (1991).

Definition 1. A partition $(\mathcal{F}, \mathcal{T})$ of \mathcal{X} into two subspaces $\mathcal{F}, \mathcal{T} \subset \mathcal{X}$ is defined as

$$A\mathcal{T} \subset E\mathcal{T} + \mathcal{B} \quad (2.5)$$

$$E\mathcal{F} \subset A\mathcal{F} + \mathcal{B} \quad (2.6)$$

$$\mathcal{X} = \mathcal{F} \oplus \mathcal{T}. \quad (2.7)$$

The subspaces \mathcal{F}, \mathcal{T} are recognized as (E, A, \mathcal{B}) and (A, E, \mathcal{B}) -invariant subspaces respectively. We shall call \mathcal{X} an invariant-partitioned subspace and denote it as $\mathcal{X} = \mathcal{F} \odot \mathcal{T}$. The notion of partition has been also

considered by Karcıas and Kalogeropoulos (1987) from a different point of view, namely from the structure of the restricted pencil in terms of its Kronecker invariants. It has been shown by Karcıas and Kalogeropoulos (1987) that every subspace \mathcal{X} is an invariant-partition if the pencil $[sE - AB]$ restricted to \mathcal{X} has no nonzero row minimal indices and B is full column rank. We point out that invariant subspaces are a special case of invariant-partitioned subspaces. Specifically, consider that an (A, E, \mathcal{B}) -invariant [resp. (E, A, \mathcal{B}) -invariant] subspaces can be recovered by Definition 1 if we set $\mathcal{T} = 0$ (resp. $\mathcal{F} = 0$).

Definition 2. A partition $(\mathcal{F}, \mathcal{T})$ is said to be *regular* if it satisfies two additional properties

$$\dim E\mathcal{T} = \dim \mathcal{T} \quad (2.8)$$

$$\dim A\mathcal{F} = \dim \mathcal{F}. \quad (2.9)$$

We shall call a subspace $\mathcal{W} \subset \mathcal{X}$ *E-regular* if (2.8) holds for \mathcal{W} and *A-regular* if (2.9) holds for \mathcal{W} .

Choosing T and S as bases for \mathcal{T} and \mathcal{F} respectively, (2.5) and (2.6) are equivalent respectively to

$$AT = ETF_T - BG_T \quad (2.10)$$

$$ES = ASF_S - BG_S \quad (2.11)$$

for some F_T, G_T, F_S, G_S . We have $T \in \mathbb{R}^{n \times \tau}$, where $\tau = \dim \mathcal{T}$, and $S \in \mathbb{R}^{n \times \sigma}$, where $\sigma = \dim \mathcal{F}$. These are the *generalized Lyapunov or Sylvester* equations. Another well-known definition (Van Dooren, 1981) is the following

Definition 3. \mathcal{F} is an (E, A) -deflating subspace if

$$\dim(E\mathcal{F} + A\mathcal{F}) = \dim \mathcal{F}. \quad (2.12)$$

We introduce now the following, which extends this definition to the case where $B \neq 0$.

Definition 4. \mathcal{F} is an (E, A, \mathcal{B}) -deflating subspace if

$$\dim(E\mathcal{F} + A\mathcal{F} + B\mathcal{X}) = \dim \mathcal{F}, \quad (2.13)$$

for some \mathcal{X} . ■

Having made these definitions, we can start our exploration of the quotient subspaces. In our approach we will need to extend some of the results in (Wonham, 1979) to the case of two maps acting simultaneously on a subspace.

If $\mathcal{F} \subset \mathcal{X}$, $x \in \mathcal{X}$, define the equivalence class $\bar{x} = \{y \in \mathcal{X} : x - y \in \mathcal{F}\}$ and the *quotient* space

\mathcal{X}/\mathcal{S} as the set of all \bar{x} . Then the canonical projection $P: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{S}$ is defined by $Px = \bar{x}$. We may also write \bar{x} as $x + \mathcal{S}$. See Wonham (1979).

In the next results, we are motivated by the fact that the equivalence of matrix pencils is defined in terms of two constant nonsingular maps, one acting on the domain and one acting on the codomain (Lewis *et al.*, 1989) [e.g. $Q(sE - A)P$]. These ancillary results will give us the first light in our exploration.

Lemma 1. Let $\mathcal{S} \subset \mathcal{X}$ and $E, A: \mathcal{X} \rightarrow \mathcal{X}$, with $sE - A$ not necessarily regular. Define $\tilde{\mathcal{S}} = A\mathcal{S} + E\mathcal{S} + B\mathcal{H}_r$ for some \mathcal{H}_r , and let P_T, Q_T be the canonical projections $P_T: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{S}$ and $Q_T: \mathcal{X} \rightarrow \mathcal{X}/\tilde{\mathcal{S}}$. Then there exist unique maps $E_T, A_T: \mathcal{X}/\mathcal{S} \rightarrow \mathcal{X}/\tilde{\mathcal{S}}$ such that

$$E_T P_T = Q_T E \quad (2.14)$$

$$A_T P_T = Q_T A. \quad (2.15)$$

Proof. Let \mathcal{T} be any complement of \mathcal{X}/\mathcal{S} so that $\mathcal{X} = \mathcal{T} \oplus \mathcal{S}$. Choose $\{t_i\}_{i=1}^r$ to be a basis for \mathcal{T} . Then, if $\bar{t}_i = P_T t_i$, a basis for \mathcal{X}/\mathcal{S} is $\{\bar{t}_i\}_{i=1}^r$, where $r = \dim \mathcal{T}$. Define E_T and A_T by

$$E_T \bar{t}_i = Q_T E t_i \quad (2.16)$$

$$A_T \bar{t}_i = Q_T A t_i. \quad (2.17)$$

To show that E_T is well-defined, suppose $\bar{x}, \bar{x}_2 \in \mathcal{X}/\mathcal{S}$ with $\bar{x}_1 = \bar{x}_2$. Then $\bar{x}_1 = x_1 + \mathcal{S}$ and $\bar{x}_2 = x_2 + \mathcal{S}$ for some $x_i \in \mathcal{T}$, and $x_1 + \mathcal{S} = x_2 + \mathcal{S}$, or $x_1 - x_2 \in \mathcal{S}$. Thus $Q_T E(x_1 - x_2) \in Q_T E\mathcal{S} \subset Q_T \tilde{\mathcal{S}} = 0$. Therefore $Q_T E x_1 = Q_T E x_2$.

Now let $x \in \mathcal{X}/\mathcal{S}$ and $x = t + s$ with $t \in \mathcal{T}$, $s \in \mathcal{S}$. Then $Q_T E x = Q_T E(t + s)$. But $Q_T E\mathcal{S} \subset Q_T \tilde{\mathcal{S}} = 0$, therefore by (2.16) $Q_T E x = Q_T E t = E_T \bar{t} = E_T P_T(t + s)$ which verifies (2.14).

Similarly we can prove that A_T is well-defined.

The conditions for the existence of the induced maps E_T and A_T in Lemma 1 are surprisingly mild; indeed, they always exist since $E\mathcal{S} \subset \tilde{\mathcal{S}}$ and $A\mathcal{S} \subset \tilde{\mathcal{S}}$.

In our approach we seek an appropriate relation between partitioned subspaces and quotient subspaces. This relationship will be the most significant result that will inspire us to utilize the notions of quotient and partitioned subspaces. The following theorem exhibits this relationship and offers the first clue in our exploration. The proof of the theorem follows easily by Lemma 1.

Theorem 1. Let $\mathcal{X} = \mathcal{S} \oplus \mathcal{T}$ and $E, A: \mathcal{X} \rightarrow \mathcal{X}$, with $sE - A$ not necessarily regular. Define $\tilde{\mathcal{S}} = A\mathcal{S} + E\mathcal{S} + B\mathcal{H}_r$ and $\tilde{\mathcal{T}} = A\mathcal{T} + E\mathcal{T} + B\mathcal{H}_r$ for some $\mathcal{H}_r, \mathcal{H}_r$. Now let P_T, P_S, Q_T, Q_S

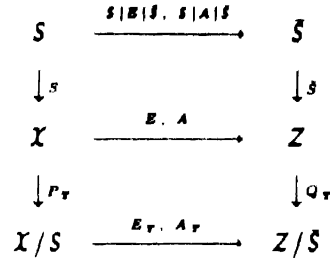


FIG. 1. Commutative diagram showing relations in Theorem 1 and equations (2.18) and (2.19).

be the canonical projections $P_T: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{S}$, $P_S: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{T}$, $Q_T: \mathcal{X} \rightarrow \mathcal{X}/\tilde{\mathcal{S}}$, $Q_S: \mathcal{X} \rightarrow \mathcal{X}/\tilde{\mathcal{T}}$. Then there exist unique maps $E_T, A_T: \mathcal{X}/\mathcal{S} \rightarrow \mathcal{X}/\tilde{\mathcal{S}}$ and $E_S, A_S: \mathcal{X}/\mathcal{T} \rightarrow \mathcal{X}/\tilde{\mathcal{T}}$ such that

$$E_T P_T = Q_T E \quad (2.18)$$

$$A_T P_T = Q_T A \quad (2.19)$$

$$E_S P_S = Q_S E \quad (2.20)$$

$$A_S P_S = Q_S A. \quad (2.21)$$

All these relations are given in the commutative diagrams of Figs 1 and 2. ■

We note that \mathcal{X}/\mathcal{S} is isomorphic to \mathcal{T} ($\mathcal{X}/\mathcal{S} \cong \mathcal{T}$) and \mathcal{X}/\mathcal{T} is isomorphic to \mathcal{S} ($\mathcal{X}/\mathcal{T} \cong \mathcal{S}$). The next step is to examine whether the subspaces $\mathcal{S}, \mathcal{T} \subset \mathcal{X}$ preserve their invariance properties under the action of the induced maps (2.18–2.21) defined in Theorem 1. The next two lemmas show that this is indeed the case. This is important because, as will be shown in Section 3, these subspaces preserve their invariance property under the action of the induced maps. This observation is the first hint toward our attempt to exploit this property for the spectrum assignability problem, which is discussed in the next section.

Lemma 2. Let $\mathcal{X} = \mathcal{S} \oplus \mathcal{T}$ and P_T, Q_T be the canonical projections $P_T: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{S}$ and $Q_T: \mathcal{X} \rightarrow \mathcal{X}/\tilde{\mathcal{S}}$. Then \mathcal{T} is an $(A_T, E_T, \mathcal{B}_T)$ -invariant subspace, where A_T, E_T are the induced maps as have been defined in Theorem 1, and $\mathcal{B}_T = Q_T B$.

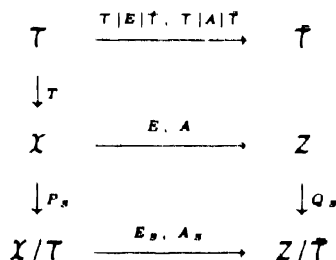


FIG. 2. Commutative diagram showing relations in Theorem 1 and equations (2.20) and (2.21).

Proof. Since \mathcal{F} is an (A, E, \mathcal{B}) -invariant subspace then the following are true

$$A\mathcal{F} \subset E\mathcal{F} + \mathcal{B} \quad (2.22)$$

$$Q_T A\mathcal{F} \subset Q_T E\mathcal{F} + Q_T \mathcal{B} \quad (2.23)$$

$$A_T P_T \mathcal{F} \subset E_T P_T \mathcal{F} + \mathcal{B}_T \quad (2.24)$$

$$A_T \mathcal{F} \subset E_T \mathcal{F} + \mathcal{B}_T \quad (2.25)$$

since $P_T: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F} \cong \mathcal{F}$. Note that (2.24) and (2.25) differ only in the basis representation of \mathcal{F} . ■

Lemma 3. Let $\mathcal{X} = \mathcal{F} \oplus \mathcal{F}$ and P_S, Q_S be the canonical projections $P_S: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F}$ and $Q_S: \mathcal{F} \rightarrow \mathcal{F}/\mathcal{F}$. Then \mathcal{F} is an $(E_S, A_S, \mathcal{B}_S)$ -invariant subspace, where A_S, E_S are the induced maps as have been defined in Theorem 1 and $\mathcal{B}_S = Q_S \mathcal{B}$. ■

In terms of these conditions the generalized Lyapunov equations for the induced maps can be written in the following form

$$A_T P_T T = E_T P_T T F_T - B_T G_T \quad (2.26)$$

$$E_S P_S S = A_S P_S S F_S - B_S G_S. \quad (2.27)$$

We can also equivalently write (2.26) and (2.27) in the following form

$$A_T T_r = E_T T_r F_T - B_T G_T \quad (2.28)$$

$$E_S S_o = A_S S_o F_S - B_S G_S. \quad (2.29)$$

Observe that these two different bases of \mathcal{F} and \mathcal{F} are related as follows

$$T_r = P_T T \quad (2.30)$$

$$S_o = P_S S, \quad (2.31)$$

where P_T and P_S have full row rank τ and σ respectively.

At this point we have to bring up some remarks for the generalized Lyapunov equations (2.10), (2.11) and (2.28), (2.29). Our interest in these equations arises from the fact that they will be the key for spectrum assignability with PD feedback. This becomes clear in the proof of Theorem 2, while in Section 3 it is evident. The first pair of equations (2.10), (2.11) imposes another condition, that is $\mathcal{F} \cap \mathcal{F} = 0$, which follows from the formulation of the problem. That is, the solutions to (2.10) and (2.11) are coupled. On the contrary, in the second pair of equations, (2.28) and (2.29) this condition is not imposed. For this reason (2.10) and (2.11) will be called the *coupled* generalized Lyapunov equations, while (2.28) and (2.29) will be called the *uncoupled* generalized Lyapunov equations. These remarks close the first formulation of the notions for quotient and partitioned subspaces. The second one is given in the sequel.

2.2. A feedback-free approach for quotient and partitioned subspaces

Let N be the left annihilator of B , and B^* be the Moore–Penrose inverse of B such that $B^* B = I_m$. Then (2.1) can be written as

$$NE\dot{x} = N A x \quad (2.32a)$$

$$u(t) = B^*(E\dot{x} - A x). \quad (2.32b)$$

Therefore the solutions of (2.1) are characterized by the solutions of (2.32a), since the solution $u(t)$ to (2.32b) always exists. We call (2.32) a *feedback-free* description of (2.1) since the solutions of (2.32a) are independent of B . This formulation will be one of our interests in the paper and will be discussed in Section 4. In order to study the feedback-free description of (2.1), we will need some preliminary results which are presented in the sequel.

Define $L = NE$ and $M = NA$, then we can extend our results for the map $sE - A$ to the map $sL - M$ as follows. This extension will later lead us to characterizations of feedback-free partitions which are of great importance.

Lemma 4. Let $\mathcal{X} = \mathcal{F} \oplus \mathcal{F}$ and $L, M: \mathcal{X} \rightarrow \mathcal{V}$. Define $\tilde{\mathcal{F}} = L\mathcal{F} + M\mathcal{F}$, and $\tilde{\mathcal{F}} = L\mathcal{F} + M\mathcal{F}$. Let $P_T, P_S, \tilde{Q}_T, \tilde{Q}_S$ be the canonical projections $P_T: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F}$, $P_S: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F}$, $\tilde{Q}_T: \mathcal{V} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$, $\tilde{Q}_S: \mathcal{V} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$. Then there exist unique maps $L_T, M_T: \mathcal{X}/\mathcal{F} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$ and $L_S, M_S: \mathcal{X}/\mathcal{F} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$ such that

$$L_T P_T = \tilde{Q}_T L \quad (2.33)$$

$$M_T P_T = \tilde{Q}_T M \quad (2.34)$$

$$L_S P_S = \tilde{Q}_S L \quad (2.35)$$

$$M_S P_S = \tilde{Q}_S M. \quad (2.36)$$

These results correspond to the commutative diagrams in Fig. 3. ■

At this point we note that for the maps L, M the equivalent invariance properties for \mathcal{F} and \mathcal{F} have been transformed to (M, L) - and (L, M) -invariance respectively. These properties are invariant under the action of the induced maps defined in Lemma 4, as is evident from the following results. The proof follows easily and therefore is omitted.

Corollary 1. Let $\mathcal{X} = \mathcal{F} \oplus \mathcal{F}$ and P_T, \tilde{Q}_T be the canonical projections $P_T: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F}$ and $\tilde{Q}_T: \mathcal{V} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$. Then \mathcal{F} is an (M_T, L_T) -invariant subspace, where M_T, L_T are the induced maps as have been defined in Lemma 4. ■

Corollary 2. Let $\mathcal{X} = \mathcal{F} \oplus \mathcal{F}$ and P_S, \tilde{Q}_S be the canonical projections $P_S: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F}$ and $\tilde{Q}_S: \mathcal{V} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$. Then \mathcal{F} is an (L_S, M_S) -invariant subspace,

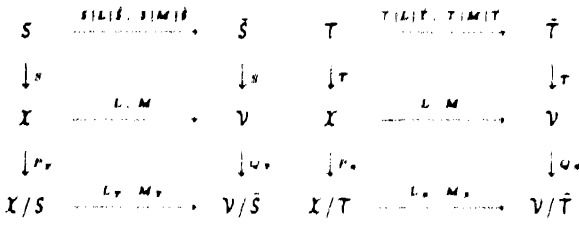


FIG. 3. Commutative diagram showing relations in Lemma 4

where L_S, M_S , are the induced maps as have been defined in Lemma 4. ■

In order to complete the relation between all these subspaces we seek a well-defined map $N_T: \mathcal{X}/\tilde{\mathcal{F}} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$. The next lemma defines such a map.

Lemma 5. Let $\mathcal{X} = \mathcal{F} \odot \tilde{\mathcal{F}}$ and $N: \mathcal{F} \rightarrow \mathcal{V}$. Let also Q_T, \tilde{Q}_T be the canonical projections $Q_T: \mathcal{F} \rightarrow \mathcal{F}/\tilde{\mathcal{F}}$ and $\tilde{Q}_T: \mathcal{V} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$. Then $N_T: \mathcal{F}/\tilde{\mathcal{F}} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$ is well-defined, that is

$$\tilde{Q}_T N = N_T Q_T. \quad (2.37)$$

This relation is shown in the commutative diagram in Fig. 4.

Proof. Let $\tilde{\mathcal{F}}$ be any complement of $\mathcal{F}/\tilde{\mathcal{F}}$ so that $\mathcal{F} = \mathcal{F}/\tilde{\mathcal{F}} \oplus \tilde{\mathcal{F}}$. Choose $\{t_i\}_{i=1}^r$ to be a basis for $\tilde{\mathcal{F}}$. Then if, $\tilde{t}_i = P t_i$, a basis for $\mathcal{F}/\tilde{\mathcal{F}}$ is $\{t_i\}_{i=1}^r$, where $r = \dim \tilde{\mathcal{F}}$, define N_T by

$$N_T \tilde{t}_i = Q_T N t_i. \quad (2.38)$$

To show that N_T is well-defined, suppose $\tilde{x}_1, \tilde{x}_2 \in \mathcal{F}/\tilde{\mathcal{F}}$ with $\tilde{x}_1 = \tilde{x}_2$. Then $\tilde{x}_1 = x_1 + \tilde{\mathcal{F}}$ and $\tilde{x}_2 = x_2 + \tilde{\mathcal{F}}$ for some $x_1 \in \mathcal{F}$, and $x_1 + \tilde{\mathcal{F}} = x_2 + \tilde{\mathcal{F}}$, or $x_1 - x_2 \in \tilde{\mathcal{F}}$. Thus $\tilde{Q}_T N(x_1 - x_2) \in \tilde{Q}_T N \tilde{\mathcal{F}} \subset \tilde{Q}_T \tilde{\mathcal{F}} = 0$. Therefore $\tilde{Q}_T N x_1 = \tilde{Q}_T N x_2$.

Now let $x \in \mathcal{F}/\tilde{\mathcal{F}}$ and $x = t + s$ with $t \in \tilde{\mathcal{F}}$, $s \in \mathcal{F}$. Then $\tilde{Q}_T N x = \tilde{Q}_T N(t + s)$. But $\tilde{Q}_T N \tilde{\mathcal{F}} \in \tilde{Q}_T \tilde{\mathcal{F}} = 0$, therefore by (2.38) $\tilde{Q}_T N x = \tilde{Q}_T N t = N_T \tilde{t} = N_T Q_T(t + s)$ which verifies (2.37). ■

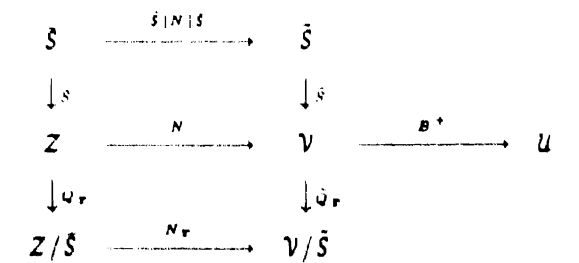


FIG. 4. Commutative diagram showing relations in Lemma 5.

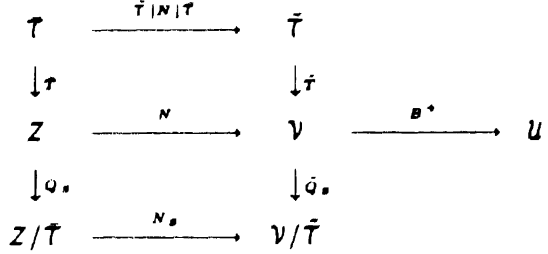


FIG. 5. Commutative diagram showing relations in Lemma 6.

The corresponding result for the connection between $\mathcal{X}/\tilde{\mathcal{F}}$ and $\mathcal{V}/\tilde{\mathcal{F}}$ is given in the following lemma.

Lemma 6. Let $\mathcal{X} = \mathcal{F} \odot \tilde{\mathcal{F}}$ and $N: \mathcal{F} \rightarrow \mathcal{V}$. Let also Q_S, \tilde{Q}_S be the canonical projections $Q_S: \mathcal{F} \rightarrow \mathcal{X}/\tilde{\mathcal{F}}$ and $\tilde{Q}_S: \mathcal{V} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$. Then $N_S: \mathcal{X}/\tilde{\mathcal{F}} \rightarrow \mathcal{V}/\tilde{\mathcal{F}}$ is well-defined that is

$$\tilde{Q}_S N = N_S Q_S. \quad (2.39)$$

This relation is shown in Fig. 5. ■

By the definitions of the induced maps the following relations hold true

$$L_T = N_T E_T \quad (2.40)$$

$$M_T = N_T A_T \quad (2.41)$$

$$L_S = N_S E_S \quad (2.42)$$

$$M_S = N_S A_S. \quad (2.43)$$

All the relations that have been defined in this section are summarized in the commutative diagram Fig. 6 for $\mathcal{X}/\tilde{\mathcal{F}}$ and in Fig. 7 for $\mathcal{X}/\tilde{\mathcal{F}}$.

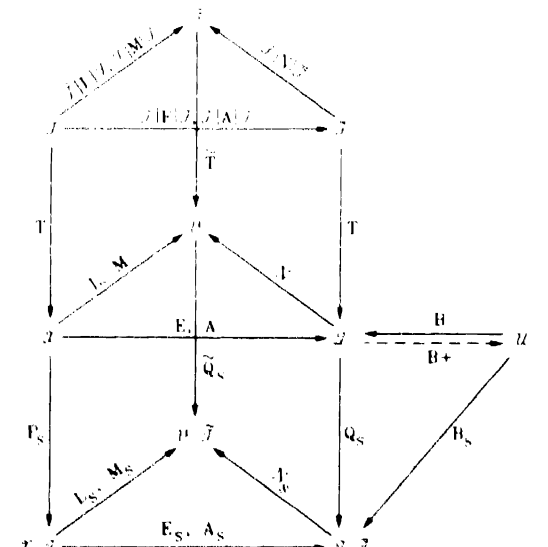


FIG. 6. Commutative diagram summarizing all relations for $\mathcal{X}/\tilde{\mathcal{F}}$.

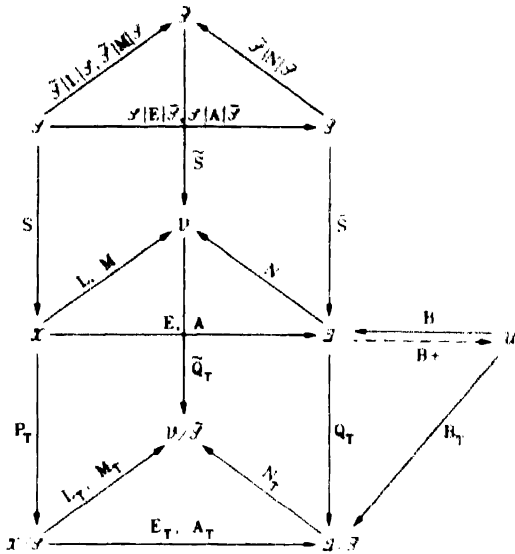


Fig. 7. Commutative diagram summarizing all relations for \mathcal{X}/\mathcal{J} .

2.3. Partitions, induced partitions and their properties

Since the partitioned subspaces are of great importance in our theory, along with the quotient subspaces which are the key tools in this geometric approach, we point out some interesting properties that a partitioned subspace enjoys. Moreover these results will bring together the quotient and the partitioned subspaces. In this attempt we will need some new definitions.

Define \mathcal{X} as an *induced mapping partition* if it satisfies

$$A_T \mathcal{F} \subset E_T \mathcal{F} + \mathcal{B}_T \quad (2.44a)$$

$$E_S \mathcal{F} \subset A_S \mathcal{F} + \mathcal{B}_S \quad (2.44b)$$

$$\mathcal{X} = \mathcal{F} \oplus \mathcal{J}, \quad (2.45)$$

$$E_T \Phi = \Psi E_S, \quad (2.46a)$$

$$A_T \Phi = \Psi A_S, \quad (2.46b)$$

where A_T, E_T, A_S, E_S are the induced maps as defined in Theorem 1. Moreover, as it has been shown by Syrmos and Lewis (1991), (2.46) guarantees that the structure of the Kronecker invariants is not destroyed under the action of this decomposition.

Similarly define \mathcal{X} as a *feedback-free induced map partition* if it satisfies

$$M_T \mathcal{F} \subset L_T \mathcal{F} \quad (2.47a)$$

$$L_S \mathcal{F} \subset M_S \mathcal{F} \quad (2.47b)$$

$$\mathcal{X} = \mathcal{F} \oplus \mathcal{J}, \quad (2.48)$$

$$L_T \Phi = \Psi L_S, \quad (2.49a)$$

$$M_T \Phi = \Psi M_S, \quad (2.49b)$$

where M_T, L_T, M_S, L_S are the induced maps as

defined in Lemma 4. Similarly (2.49) guarantees the Kronecker structure of the decomposition.

The next theorem exhibits some of the properties that these partitions enjoy. Actually, we will concentrate on the induced mapping partition since it is the one that we will use more frequently later on. Note that the next theorem holds for each one of these three partitions; that is the partitions satisfying (2.5–7), (2.44–46) and (2.47–48). We point out that this theorem through its proof indicates the use of the generalized Lyapunov equations for the spectrum assignability problem using PD feedback.

Theorem 2. $\mathcal{X} = \mathcal{F} \oplus \mathcal{J}$ is an induced mapping partition if and only if for any $x(0) \in \mathcal{X}$ there exists an input $u(t)$ such that:

- $x(t) \in \mathcal{X}$ for $t \geq 0$, for some $u(t)$.
- The Laplace transforms of $x(t)$, $u(t)$ restricted to \mathcal{F} are strictly proper.
- The Laplace transforms of $\dot{x}(t)$, $u(t)$ restricted to \mathcal{F} have no poles at the origin.

Note. For comparison of (ii) and (iii), note that a strictly proper transfer function has no poles at infinity.

Proof. (Necessity) Let S, T provide us with bases for \mathcal{F}, \mathcal{J} respectively. Then the generalized Lyapunov equations (2.10) and (2.11) after some manipulation can also be written in the form

$$(sE - A)T(sI_r - F_T)^{-1} = ET + BG_T(sI_r - F_T)^{-1} \quad (2.50)$$

$$(E - s^{-1}A)S(F_S - s^{-1}I_o)^{-1} = AS + BG_S(F_S - s^{-1}I_o)^{-1} \quad (2.51)$$

with s a complex variable. Premultiply (2.50) and (2.51) by Q_T and Q_S respectively, where Q_T and Q_S have been defined in Theorem 1. Taking into account (2.18–21), (2.50) and (2.51) become

$$(sE_T - A_T)P_T T(sI_r - F_T)^{-1} = E_T P_T T + B_T G_T(sI_r - F_T)^{-1} \quad (2.52)$$

$$(E_S - s^{-1}A_S)P_S S(F_S - s^{-1}I_o)^{-1} = A_S P_S S + B_S G_S(F_S - s^{-1}I_o)^{-1}. \quad (2.53)$$

Then by using (2.30) and (2.31) in order to change the basis representation for \mathcal{F}, \mathcal{J} , we have

$$(sE_T - A_T)T_r(sI_r - F_T)^{-1} = E_T T_r + B_T G_T(sI_r - F_T)^{-1} \quad (2.54)$$

$$(E_S - s^{-1}A_S)S_o(F_S - s^{-1}I_o)^{-1} = A_S S_o + B_S G_S(F_S - s^{-1}I_o)^{-1}. \quad (2.55)$$

Restrict (2.1) to \mathcal{T} and take the Laplace transform of the restricted version. Then (2.10) becomes

$$(sE - A)X_T(s) = Ex_T(0^-) + BU_T(s), \quad (2.56)$$

where $X_T(s)$ and $U_T(s)$ are the restrictions of $X(s)$ and $U(s)$ on \mathcal{T} with respect to (w.r.t.) T basis. Now take a Laplace transform [denoted $L(\cdot)$] of (2.1) defining

$$L(\dot{x}) = X'(s) \quad (2.57)$$

$$L(x) = s^{-1} \left(X'(s) + \int_{-\infty}^0 x(t) dt \right). \quad (2.58)$$

Then (2.1) restricted to \mathcal{S} becomes

$$(E - s^{-1}A)X'_S(s)s = Ax_S(0^-) + BU_S(s)s, \quad (2.59)$$

where $X'_S(s)$ and $U_S(s)$ are the restrictions of $X'(s)$ and $U(s)$ on S w.r.t. to S basis representation. By premultiplying now (2.56) and (2.59) by Q_T and Q_S respectively and using the same reasoning as above (2.56) and (2.57) become

$$(sE_T - A_T)X_{T_i}(s) = E_T x_{T_i}(0^-) + B_T U_T(s) \quad (2.60)$$

$$(E_S - s^{-1}A_S)X'_{S_i}(s)s = Ax_{S_i}(0^-) + B_S U_S(s) \quad (2.61)$$

where X_{T_i} , U_{T_i} , and X'_{S_i} , U_{S_i} are the restrictions on \mathcal{T} , \mathcal{S} respectively w.r.t. T_i and S_i bases characterization. By comparing (2.60), (2.61) and (2.54), (2.55) it is seen that the Laplace transforms of $x(t)$ and $u(t)$ restricted to \mathcal{T} are strictly proper and that the Laplace transforms of $\dot{x}(t)$ and $\dot{u}(t)$ restricted to \mathcal{S} have no poles at the origin. Moreover it is clear that $x(t) \in \mathcal{X}$ for $t \geq 0$.

(Sufficiency) If for every $x_{T_i}(0^-) \in \mathcal{T}$ there are strictly proper rational $u(t)$ and $x(t)$ restricted to \mathcal{T} that satisfy (2.1) and $x_{T_i}(t) = P_T x_T(t) = P_T T x(t) \in \mathcal{T}$, then

$$X_{T_i} = s^{-1} \sum_{i=0}^{\infty} x'_{T_i} s^{-i}, \quad U_{T_i} = s^{-1} \sum_{i=0}^{\infty} u'_{T_i} s^{-i} \quad (2.62)$$

with all $v'_{T_i} \in \mathcal{T}$, and (2.60) shows that

$$E_T v_{T_i}^0 = E_T v_{T_i}(0^-) \quad (2.63a)$$

$$E_T v_{T_i}^1 = A_T v_{T_i}^0 + B_T u_{T_i}^0. \quad (2.63b)$$

If T_i is a basis for \mathcal{T} , then (2.63a) shows that $E_T T_{i0} = E_T T_i$ and (2.63b) shows that

$$\begin{aligned} A_T T_{i0} &= E_T T_i F_T - B_T G_T \\ &= E_T T_{i0} - B_T G_T \end{aligned} \quad (2.64)$$

for some F_T and G_T . According to (2.28) this identifies \mathcal{T} as an $(A_T, E_T, \mathcal{B}_T)$ -invariant subspace.

Similarly we can prove that \mathcal{S} is an $(E_S, A_S, \mathcal{B}_S)$ -invariant subspace, which completes the proof.

Concluding the proof we address the interest of the reader to equations (2.54) and (2.55), where the terms $(sI_r - F_T)$ and $(F_S - s^{-1}I_\sigma)$ are presented. As we will show in Section 3 the spectrum assignability depends on the selection of F_S and F_T in (2.28) and (2.29). ■

A problem that will arise in the next section is the regularity of the closed-loop system under the action of PD feedback. The partitioned subspaces as well as the quotient subspaces are closely related to regularity, as we shall see. The next theorem will provide us with the tools to confront the closed-loop regularity problem. Moreover, it justifies the choice of the particular quotient subspaces which we have used in this section.

Theorem 3. Let $\mathcal{X} = \mathcal{T} \oplus \mathcal{S}$. Then the following are equivalent:

- (i) \mathcal{S} , \mathcal{T} is a regular partition.
- (ii)

$$\dim(E\mathcal{T} + A\mathcal{T} + \mathcal{B}) = \dim \mathcal{T} + \dim \mathcal{B}_{T_i} \quad (2.65)$$

$$\dim(E\mathcal{S} + A\mathcal{S} + \mathcal{B}) = \dim \mathcal{S} + \dim \mathcal{B}_{S_i} \quad (2.66)$$

for some \mathcal{B}_{T_i} , \mathcal{B}_{S_i} given by

$$\mathcal{B}_{T_i} \subset E\mathcal{T} \oplus \mathcal{B}_{T_i} \quad (2.67)$$

$$\mathcal{B}_{S_i} \subset A\mathcal{S} \oplus \mathcal{B}_{S_i} \quad (2.68)$$

$$\mathcal{B} = \mathcal{B}_{T_i} \oplus \mathcal{B}_{T_j} \quad (2.69)$$

$$\mathcal{B} = \mathcal{B}_{S_i} \oplus \mathcal{B}_{S_j}. \quad (2.70)$$

Before giving the proof of the theorem let us comment on condition (ii). This is a technical condition that seems cumbersome at first glance. As we shall see in the next section, condition (ii) is the key to relating the open-loop properties we are discussing here to the properties of the closed-loop system.

Proof. (i) \Rightarrow (ii): Assume that \mathcal{X} , \mathcal{T} is a regular partition then

$$\dim E\mathcal{T} = \dim \mathcal{T} \quad (2.71)$$

$$\dim A\mathcal{S} = \dim \mathcal{S}. \quad (2.72)$$

But also since \mathcal{X} is a partition we have

$$A\mathcal{T} + \mathcal{B}_{T_i} \subset E\mathcal{T} \oplus \mathcal{B}_{T_i} \quad (2.73)$$

$$E\mathcal{S} + \mathcal{B}_{S_i} \subset A\mathcal{S} \oplus \mathcal{B}_{S_i} \quad (2.74)$$

Then the following holds true:

$$\begin{aligned} \dim((A\mathcal{F} + \mathcal{B}_{T_1}) \cap (E\mathcal{F} \oplus \mathcal{B}_{T_1})) \\ = \dim(A\mathcal{F} + \mathcal{B}_{T_1}) \end{aligned} \quad (2.75)$$

$$\begin{aligned} \dim((E\mathcal{F} + \mathcal{B}_{S_1}) \cap (A\mathcal{F} \oplus \mathcal{B}_{S_1})) \\ = \dim(E\mathcal{F} + \mathcal{B}_{S_1}). \end{aligned} \quad (2.76)$$

Taking into account (2.71) and (2.72) then (2.75) and (2.76) become

$$\begin{aligned} \dim(A\mathcal{F} + \mathcal{B}_{T_1}) + \dim(E\mathcal{F} \oplus \mathcal{B}_{T_1}) \\ - \dim(E\mathcal{F} + A\mathcal{F} + \mathcal{B}) = \dim(A\mathcal{F} + \mathcal{B}_{T_1}) \end{aligned}$$

$$\begin{aligned} \dim(E\mathcal{F} + \mathcal{B}_{S_1}) + \dim(A\mathcal{F} \oplus \mathcal{B}_{S_1}) \\ - \dim(E\mathcal{F} + A\mathcal{F} + \mathcal{B}) = \dim(E\mathcal{F} + \mathcal{B}_{S_1}). \end{aligned}$$

or

$$\dim \mathcal{F} + \dim \mathcal{B}_{T_1} = \dim(E\mathcal{F} + A\mathcal{F} + \mathcal{B})$$

$$\dim \mathcal{F} + \dim \mathcal{B}_{S_1} = \dim(E\mathcal{F} + A\mathcal{F} + \mathcal{B}).$$

(i) \Leftarrow (ii): Straightforward. ■

Theorem 3 and specifically condition (ii) will be of significance in our approach for the regularity of the closed-loop system under the action of PD feedback. Moreover, condition (ii) is the key for the construction of induced mapping partitions that preserve the property of regularity.

3. PARTITIONED SUBSPACES, QUOTIENT SUBSPACES AND PROPORTIONAL-PLUS-DERIVATIVE FEEDBACK

In this section we show how the partitioned subspaces, which constitute open-loop characterizations, give information about the closed-loop system under proportional-plus-derivative (PD) feedback. We shall study how we can transform these partitioned subspaces to induced mapping partitioned subspaces using the notions of the quotient subspaces. Moreover, we guarantee the regularity of the closed-loop system on these subspaces. We show how to construct induced mapping partitions which preserve the closed-loop regularity of the system. As a result, we shall be able to solve Lyapunov equations that are equivalent to those of (2.10) and (2.11). Specifically Theorem 7 designates a computationally stable method for the calculation of \mathcal{H}_S and \mathcal{H}_T , that guarantees this equivalence. This computation is done in the very beginning of our design technique for spectrum assignability, which is based on the solution of the uncoupled generalized Lyapunov equations. Moreover we present a technique that constructs PD feedbacks for the desired closed-loop behavior of (2.1). This design technique is based on the generalized Hessenberg form. Therefore it is easily implementable

and computationally stable. At the end of this section, Lemma 7 shows that the closed-loop spectrum assigned by the solutions of the uncoupled generalized Lyapunov equations is the same as those of the coupled ones. In order to achieve this goal we first present some preliminary results which will help us in our endeavor.

3.1. Partitions and PD feedback

If the PD feedback

$$u = -K_2\dot{x} + K_1x \quad (3.1)$$

is applied to (2.1), the resulting closed-loop system is

$$(E + BK_2)\dot{x} = (A + BK_1)x. \quad (3.2)$$

Even if $E = I$, (3.2) may not be a state-variable system since $(E + BK_2)$ is generally singular. Moreover in the case where $E + BK_2$ is ill conditioned it is wise to avoid its inversion and study the closed-loop system as a singular system.

The next result shows how to use the Lyapunov equations (2.10), (2.11) or the Lyapunov equations (2.28), (2.29) to construct a PD feedback that provides a certain closed-loop invariance property.

Theorem 4. $\mathcal{X} = \mathcal{F} \oplus \mathcal{T}$ is a partitioned subspace, or equivalently an induced mapping partitioned subspace, if and only if there exist K_1, K_2 , such that

$$(A + BK_1)\mathcal{F} \subset E\mathcal{F} \quad (3.3)$$

$$(E + BK_2)\mathcal{T} \subset A\mathcal{T}, \quad (3.4)$$

or equivalently

$$(A_T + B_T K_T)\mathcal{T} \subset E_T \mathcal{T} \quad (3.5)$$

$$(E_S + B_S K_S)\mathcal{F} \subset A_S \mathcal{F}. \quad (3.6)$$

Proof. Let S, T be bases for \mathcal{F}, \mathcal{T} respectively and F_S, G_S, F_T, G_T satisfy (2.10) and (2.11). Select the feedback gains as

$$[G_T \ 0] = K_1[T \ S] \quad (3.7)$$

$$[0 \ G_S] = K_2[T \ S] \quad (3.8)$$

so that $K_1 S = K_2 T = 0$. Then (2.50) and (2.51) become

$$(sE - (A + BK_1))T(sI_r - F_T)^{-1} = ET \quad (3.9)$$

$$((E + BK_2) - s^{-1}A)S(F_S - s^{-1}I_o)^{-1} = AS. \quad (3.10)$$

Also equivalent to (2.50) and (2.51) are (2.54)

and (2.55) which can be written as

$$(sE_T - (A_T + B_T K_T))T_\tau (sI - F_\tau)^{-1} = E_T T_\tau \quad (3.11)$$

$$((E_S + B_S K_S) - s^{-1}A_S)S_\sigma (F_\sigma - s^{-1}I_\sigma)^{-1} = A_S S_\sigma \quad (3.12)$$

where T_τ and S_σ is another basis representation for \mathcal{T} and \mathcal{S} respectively. Note that $K_T = K_1 P_T^+$, $K_S = K_2 P_S^+$, where P_T^+ , P_S^+ are the Moore-Penrose inverses of P_T , P_S respectively.

Thus (3.9-12) along with Theorem 2 guarantee (3.3-6). The converse argument follows easily. ■

We emphasize that the proof of this theorem provides an extremely convenient design technique for PD feedback, since it relies on the solutions of the generalized Lyapunov equations. Moreover, the terms $(sI_\tau - F_\tau)$ and $(F_S - s^{-1}I_\sigma)$ in (3.11) and (3.12) exhibit the importance of F_T and F_S in (2.28) and (2.29) for the spectrum assignability problem. These terms assign the closed-loop trajectories of (3.2) on the subspaces \mathcal{S} and \mathcal{T} . Finally, Theorem 4 shows the relation between different representations, that is between (2.10)/(2.11) and (2.28)/(2.29), which we shall soon discuss further.

Note that we make no claim on the regularity of the closed-loop system (3.2). In fact, the next results show that certain conditions must hold in order to ensure the regularity of the closed-loop system (3.2) on \mathcal{X} .

Theorem 5. Let $\mathcal{X} = \mathcal{S} \oplus \mathcal{T}$ be a partitioned subspace. Define any K_1 , K_2 such that $K_1 S = K_2 T = 0$. Then the pencil $[s(E + BK_2) - (A + BK_1)]$ is of full rank restricted to \mathcal{X} if and only if $[s(E + BK_2) - A]$ and $[sE - (A + BK_1)]$ are of full rank restricted to \mathcal{S} and \mathcal{T} respectively.

Proof. Consider $\{t_i\}_{i=1}^r$ and $\{s_i\}_{i=1}^\sigma$ as the bases for \mathcal{T} and \mathcal{S} respectively. Since $\mathcal{X} = \mathcal{S} \oplus \mathcal{T}$ then we can choose a basis for \mathcal{X} the set $x_i = \{t_1, \dots, t_r, s_1, \dots, s_\sigma\}$, $i \in \{1, \dots, \chi\}$ where $\chi = \dim \mathcal{X}$.

If the pencil $[s(E + BK_2) - (A + BK_1)]$ is of full rank on \mathcal{X} , then choose a basis for \mathcal{X} as above and select $s \in \mathbb{C} - [s(E + BK_2), (A + BK_1)]$, then we get

$$[s(E + BK_2) - (A + BK_1)]x_i = w_i \quad (3.13)$$

where w_i are linearly independent. Let us now impose the condition $K_1 S = K_2 T = 0$, keeping in mind the structure of the basis for \mathcal{X} clearly

(3.13) becomes

$$[s(E + BK_2) - A]s_i = w_{i+\tau}, \quad i \in \{1, \dots, \sigma\} \quad (3.14)$$

$$[sE - (A + BK_1)]t_i = w_i, \quad i \in \{1, \dots, \tau\} \quad (3.15)$$

where the sets $\{w_{i+\tau}\}_{i=1}^\sigma$ and $\{w_i\}_{i=1}^\tau$ consist of linearly independent vectors.

Conversely using (3.14) and (3.15) and taking into account that $\mathcal{S} \cap \mathcal{T} = 0$, that is the sets $\{w_{i+\tau}\}_{i=1}^\sigma$ and $\{w_i\}_{i=1}^\tau$ span two different subspaces that have the property $\mathcal{S} \cap \mathcal{T} = 0$. ■

From the above theorem we are led directly to the next theorem which displays the conditions for regularity of the closed-loop system.

Theorem 6. Let $\mathcal{X} = \mathcal{S} \oplus \mathcal{T}$ be a partitioned subspace and K_1 , K_2 satisfy (3.5) and (3.6) with $K_1 S = K_2 T = 0$. Then the closed-loop system is regular if and only if $(\mathcal{S}, \mathcal{T})$ is a regular partition.

Proof. Note that from Theorem 5, it is equivalent to show that $[s(E + BK_2) - A]$ and $[sE - (A + BK_1)]$ are of full rank restricted to \mathcal{S} and \mathcal{T} respectively. Equations (2.10) and (2.11) show the existence of a solution for every $x(0) \in \mathcal{X}$. This is independent of G_S , G_T and it is unique if and only if there are unique solutions to (2.10) and (2.11) for F_S , F_T . That is ET and AS have full column rank or equivalently $\dim E\mathcal{T} = \dim \mathcal{T}$ and $\dim A\mathcal{S} = \dim \mathcal{S}$. Therefore \mathcal{X} is identified as a regular partition. ■

At this point, it is clear that for every partition $(\mathcal{S}, \mathcal{T})$ that satisfies one of the conditions in Theorem 3 there will always exist a PD feedback that guarantees the closed-loop regularity of the system on \mathcal{X} . Specifically, derivative feedback is used on \mathcal{S} and proportional feedback is used on \mathcal{T} .

3.2. The construction of a well-defined partition

It is important now to see how the property of regularity of the closed-loop system of a partitioned subspace is interpreted for an induced mapping partitioned subspace. This will be of significance since our final goal is to find an equivalent closed-loop system for which we can solve the generalized Lyapunov equations for the pole assignment problem. This need arises from the fact that these equations for the original closed-loop are cumbersome from the point of a stable computational solution, since they are coupled. The next theorem demonstrates under which conditions we can use equations (2.28) and (2.29) without destroying

the closed-loop loop regularity of the system as has already been defined in terms of equations (2.10) and (2.11) in Theorem 6. Specifically we seek conditions on $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{S}}$ such that the induced mapping partition is well-defined, that is, it preserves the closed-loop regularity.

Theorem 7. Define a partition $\mathcal{X} = \mathcal{S} \odot \mathcal{T}$ such that \mathcal{T} and \mathcal{S} are (E, A, B) -deflating subspaces and choose any K_1, K_2 , where $K_1 S = K_2 T = 0$. Then the pencil $[s(E + BK_2) - (A + BK_1)]$ restricted to \mathcal{X} is of full rank if and only if the pencils $[s(E_S + B_S K_S) - A_S]$ and $[sE_T - (A_T + B_T K_T)]$ restricted to \mathcal{S} and \mathcal{T} respectively are of full rank, where A_S, A_T, E_S, E_T define an induced mapping partition.

Proof. Define $\tilde{\mathcal{F}}, \tilde{\mathcal{S}}$ as

$$\tilde{\mathcal{T}} = A\mathcal{T} + E\mathcal{T} + B\mathcal{K}_T \quad (3.16)$$

$$\tilde{\mathcal{S}} = A\mathcal{S} + E\mathcal{S} + B\mathcal{K}_S \quad (3.17)$$

such that $\mathcal{T}, \mathcal{S}, \tilde{\mathcal{T}}, \tilde{\mathcal{S}}$ satisfy

$$\dim \mathcal{T} = \dim \tilde{\mathcal{T}} \quad (3.18)$$

$$\dim \mathcal{S} = \dim \tilde{\mathcal{S}} \quad (3.19)$$

Choose as in Theorem 5 $\{t_i\}_{i=1}^r, \{s_i\}_{i=1}^\sigma, \{x_i\}_{i=1}^n$ be bases for \mathcal{T}, \mathcal{S} and \mathcal{X} respectively. Since the pencil $[s(E + BK_2) - (A + BK_1)]$ is of full rank restricted to \mathcal{X} , (3.14) and (3.15) hold true. Premultiply (3.14) by Q_S and (3.15) by Q_T . Then by using also (2.18–21), (2.30) and (2.31) we get

$$[s(E_S + B_S K_S) - A_S]s_\sigma = Q_S w_{1+\sigma}, \quad i \in \{1, \dots, \sigma\} \quad (3.20)$$

$$[sE_T - (A_T + B_T K_T)]t_\tau = Q_T w_i, \quad i \in \{1, \dots, \tau\} \quad (3.21)$$

where clearly $\{s_\sigma\}_{\sigma=1}^\sigma$ and $\{t_\tau\}_{\tau=1}^\tau$ are bases for \mathcal{S} and \mathcal{T} respectively. The pencils $[s(E_S + B_S K_S) - A_S]$ and $[sE_T - (A_T + B_T K_T)]$ are of full rank restricted to \mathcal{S} and \mathcal{T} respectively if and only if $\text{rank } Q_S = \sigma$ and $\text{rank } Q_T = \tau$. But this is always true by the construction of the induced maps, that is (3.18) and (3.19) always hold. The converse follows easily by using (3.20) and (3.21) and the same reasoning as before exploiting the fact that $\mathcal{S} \cap \mathcal{T} = 0$. ■

Theorem 7 is of importance since it provides us, by construction, with regularity conditions for the closed-loop system in terms of an induced mapping partition. In order to clarify this statement note that we can construct different classes of induced mapping partition. This freedom is due to the fact that we can arbitrarily pick $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$ in Theorem 7. Therefore Theorem 7 “chooses” the class of induced mapping

partitions for which the closed-loop regularity is guaranteed. Let $\mathcal{X} = \mathcal{S} \odot \mathcal{T}$ be a partition defined as in Theorem 7. Then if there exists an induced map partition such that

$$\begin{aligned} \dim(E_T \mathcal{T} + A\mathcal{T} \mathcal{T} + B_T \mathcal{K}_T) \\ = \dim(E\mathcal{T} + A\mathcal{T} + B\mathcal{K}_T) \end{aligned} \quad (3.22)$$

$$\begin{aligned} \dim(E_S \mathcal{S} + A_S \mathcal{S} + B_S \mathcal{K}_S) \\ = \dim(E\mathcal{S} + A\mathcal{S} + B\mathcal{K}_S) \end{aligned} \quad (3.23)$$

or equivalently

$$\begin{aligned} \text{rank}(E_T T_i + A_T T_i + B_T H_i) \\ = \text{rank}(ET + A\mathcal{T} + BH_T) \end{aligned} \quad (3.24)$$

$$\begin{aligned} \text{rank}(E_S S_\sigma + A_S S_\sigma + B_S H_\sigma) \\ = \text{rank}(ES + AS + BH_S). \end{aligned} \quad (3.25)$$

\mathcal{X} will be called a well-defined partition. Thus, we can always construct an induced mapping partition that guarantees the regularity of the closed-loop system. Theorem 7 shows how to construct such an induced mapping partition.

We point out that the calculation of \mathcal{K}_S and \mathcal{K}_T is of critical importance, since it indicates the choice of $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$ for which the induced mapping partition is well-defined. We designate that the calculation of H_S and H_T has to be performed at the beginning of the proposed design technique in order to construct a well-defined partition. The construction of this partition is closely related to the choice of H_S and H_T that directly involves condition (ii) of Theorem 3. It follows that the selection of such a partition is not unique but there exists a class of well-defined partitions, that is the induced partitions that are constructed according to Theorem 7. All these notions will be made clear by presenting a computationally stable algorithm for the calculation of H_S and H_T that satisfy (3.18) and (3.19), which we now do. Once H_S and H_T have been selected the construction of the partition follows.

The construction of a well-defined induced partition is closely related to condition (ii) in Theorem 3 as we have already mentioned. This will be evident by presenting a computationally stable algorithm for the calculation of such induced partitions. The first step to this algorithm is the computation of H_S and H_T that satisfies (3.18) and (3.19).

Define a partition such that

$$\dim(A\mathcal{T} + E\mathcal{T}) \leq \dim \mathcal{T} \quad (3.26)$$

$$\dim(A\mathcal{S} + E\mathcal{S}) \leq \dim \mathcal{S}. \quad (3.27)$$

Then according to Theorem 5 there always exist H_T and H_S such that

$$\dim(A\mathcal{T} + E\mathcal{T} + B\mathcal{K}_T) = \dim \mathcal{T} \quad (3.28)$$

$$\dim(A\mathcal{S} + E\mathcal{S} + B\mathcal{K}_S) = \dim \mathcal{S}. \quad (3.29)$$

Define now column operations J_T and J_S on B so that

$$\begin{aligned} BJ_T &= B[J_{T_1} \ J_{T_2}] = [BJ_{T_1} \ BJ_{T_2}] \\ &= [B_{T_1} \ B_{T_2}] \end{aligned} \quad (3.30)$$

$$\begin{aligned} BJ_S &= B[J_{S_1} \ J_{S_2}] = [BJ_{S_1} \ BJ_{S_2}] \\ &= [B_{S_1} \ B_{S_2}], \end{aligned} \quad (3.31)$$

with B_{T_1} and B_{S_1} such that

$$A\mathcal{T} + \mathcal{B} = E\mathcal{T} \oplus \mathcal{B}_{T_1} \quad (3.32)$$

$$E\mathcal{S} + \mathcal{B} = A\mathcal{S} \oplus \mathcal{B}_{S_1} \quad (3.33)$$

which implies that

$$\mathcal{B}_{T_2} \subset E\mathcal{T} \oplus \mathcal{B}_{T_1} \quad (3.34)$$

$$\mathcal{B}_{S_2} \subset A\mathcal{S} \oplus \mathcal{B}_{S_1} \quad (3.35)$$

This can be performed in a simple and effective way by checking the column rank of B . Therefore this column operation is a simple shuffle of the columns of B , where on each step we check the rank of the specific columns. Now choose as H_T , H_S any columns of J_T , J_S such that the following conditions hold true

$$\dim(A\mathcal{T} + E\mathcal{T} + B\mathcal{H}_T) = \dim \mathcal{T} \quad (3.36)$$

$$\dim(A\mathcal{S} + E\mathcal{S} + B\mathcal{H}_S) = \dim \mathcal{S} \quad (3.37)$$

for the specific choice of \mathcal{H}_T and \mathcal{H}_S as has been defined. But (3.36) and (3.37) guarantee that for the specific selection of H_T and H_S , (3.18) and (3.19) hold. Thus this technique, which is based on row operations on B (i.e. computationally stable operations), completely defines and computes the \mathcal{H}_T and \mathcal{H}_S so that (3.36) and (3.37) hold true. Specifically, this procedure defines/computes $\tilde{\mathcal{T}}$ and $\tilde{\mathcal{S}}$, as they have been defined in Theorem 7. For these $\tilde{\mathcal{T}}$ and $\tilde{\mathcal{S}}$ which the induced partition is well-defined as follows from Theorem 7.

The algorithm for the construction/computation of a well-defined induced mapping partition is given below.

Algorithm 1

Step 1: Define a partition $(\mathcal{S}, \mathcal{T})$ such that (3.26), (3.27) hold true.

Step 2: Perform row operations on B as defined in (3.28–31) and compute H_T and H_S .

Step 3: Compute $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$ by using (3.16) and (3.17).

Step 4: For the computed $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$ calculate P_S , Q_S and P_T , Q_T as defined in Theorem 7.

Step 5: Using equations (2.18–21) compute E_S , A_S , E_T , A_T . ■

It is now clear from Theorems 1 and 7 that E_S , A_S , E_T , A_T which are computed by using the Algorithm 1 constitute a well-defined induced

mapping partition for the original $(\mathcal{S}, \mathcal{T})$ partition.

The following remark goes a long way toward showing the relation of the constructed well-defined partition and the reduced order proportional state-variable feedback. The application of a PD feedback on such a partition can be decomposed to the application of two reduced order *proportional state-variable* feedbacks one acting on $\mathcal{S} \cong \mathcal{X}/\mathcal{T}$ and the other on $\mathcal{T} \cong \mathcal{X}/\mathcal{S}$. This observation follows from condition (ii) of Theorem 3 and the notions that are presented in Lewis and Özaldıran (1989).

3.3. The spectrum assignability problem

It is clear from our discussion up to this point that the solution of the generalized Lyapunov equations is extremely important for the pole placement problem using PD feedback. Through Theorem 4 it became evident that the choice of F_S and F_T assign the closed-loop trajectories of (3.2) on the subspaces \mathcal{S} and \mathcal{T} respectively. Specifically given the desired closed-loop trajectory on these subspaces, we define F_S and F_T that give this desired behavior of the closed-loop system. Moreover once F_S and F_T are specified using the generalized Lyapunov equations we find \mathcal{S} and \mathcal{T} . On these subspaces the closed-loop spectrum is assigned. But the solution of these equations defines G_S and G_T the knowledge of which will lead us to the construction of a PD feedback that assigns the desired spectrum.

At this point we will show that solving the uncoupled generalized Lyapunov equations (2.28) and (2.29) is equivalent to solving the coupled (2.10) and (2.11). The terminology "coupled and uncoupled" has been justified in Section 2. Recall that an extra condition that is imposed on (2.10) and (2.11) is $\mathcal{S} \cap \mathcal{T} = 0$. This condition is not imposed on (2.28) and (2.29). Therefore the independent solutions of (2.28) and (2.29) will lead us to the solutions of (2.10) and (2.11).

Consider a partition and an induced mapping partition constructed as in Algorithm 1. Define the projection matrix

$$P = \begin{pmatrix} P_S \\ P_T \end{pmatrix} \quad (3.38)$$

where $P \in \mathcal{R}^{n \times n}$, $P_S \in \mathcal{R}^{n \times n}$ and $P_T \in \mathcal{R}^{r \times n}$. Then P_S and P_T have full row rank since they are canonical projections as defined in Theorem 1. Specifically rank $P_S = \sigma$ and rank $P_T = \tau$.

Consider now $X = [S \ T]$ to be a basis for $\mathcal{X} = \mathcal{S} \oplus \mathcal{T}$, then

$$PX = \begin{pmatrix} P_S X \\ P_T X \end{pmatrix} = \begin{pmatrix} P_S S & P_S T \\ P_T S & P_T T \end{pmatrix} = \begin{pmatrix} S_\sigma & 0 \\ 0 & T_\tau \end{pmatrix}. \quad (3.39)$$

But since $P_S: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F} \cong \mathcal{Y}$ and $P_T: \mathcal{X} \rightarrow \mathcal{X}/\mathcal{F} \cong \mathcal{F}$, $P_S T = 0$ and $P_T S = 0$. Hence (3.39) becomes

$$PX = (PS \quad PT) = \begin{pmatrix} S_o & 0 \\ 0 & T_r \end{pmatrix} \quad (3.40)$$

or

$$X = (S \quad T) = P^* \begin{pmatrix} S_o & 0 \\ 0 & T_r \end{pmatrix}. \quad (3.41)$$

Equation (3.43) implies that given S_o and T_r we can find the solutions $X = [S \quad T]$. Therefore solving (2.28) and (2.29) is equivalent to solving (2.10) and (2.11).

It is now evident that assigning the poles using (2.28) and (2.29) is equivalent to assigning the poles of the closed-loop system using (2.10) and (2.11). This last observation demonstrates the importance of the quotient and induced mapping partitioned subspaces.

In order to achieve our goal we have to find S_o and T_r on which the closed-loop system is regular and on which the desired poles of the system are assigned. In this attempt we will need some well-known results which are briefly stated in the sequel.

The subspaces \mathcal{Y} , \mathcal{F} of a well-defined partition are $(E_S, A_S, \mathcal{B}_S)$ - and $(A_T, E_T, \mathcal{B}_T)$ -invariant subspaces, and are closed under addition, so that each one has a largest member. We symbolize the *supremal* $(E_S, A_S, \mathcal{B}_S)$ -invariant subspace as \mathcal{Y}_o^* and the *supremal* $(A_T, E_T, \mathcal{B}_T)$ -invariant subspace as \mathcal{F}_r^* . The next theorem shows how to compute \mathcal{Y}_o^* and \mathcal{F}_r^* (Wong, 1974)

Theorem 8. Consider the following subspace recursions

$$\mathcal{X}_{k+1} = A_T^{-1}(E_T \mathcal{X}_k + \mathcal{B}_T), \quad \text{with } \mathcal{X}_0 = \mathcal{X}^r \quad (3.42)$$

$$\mathcal{Y}_{k+1} = E_S^{-1}(A_S \mathcal{Y}_k + \mathcal{B}_S), \quad \text{with } \mathcal{Y}_0 = \mathcal{Y}^o \quad (3.43)$$

then $\mathcal{Y}_o^* = \mathcal{Y}_\infty$ and $\mathcal{F}_r^* = \mathcal{X}_\infty$. ■

As an alternative to (3.42) and (3.43) \mathcal{Y}_o^* and \mathcal{F}_r^* may be computed by the numerically convenient Singular System Structure Algorithm (Lewis, 1986).

3.3.1. The induced pencils. Consider a partition $(\mathcal{Y}, \mathcal{F})$. Define the *causal* pencil of (2.1) as

$$P(s) = [sE - A \quad B], \quad (3.44)$$

and the *anticausal* pencil of (2.1) as

$$D(z) = [E - zA \quad zB], \quad (3.45)$$

where $z = s^{-1}$. Having computed a well-defined partition using Algorithm 1, that is we constructed E_S, A_S, E_T and A_T while (3.28) and (3.29) hold true, then define the induced causal

pencil of a well-defined induced partition as

$$P_m(s) = [sE_T - A_T \quad B_T], \quad (3.46)$$

and the anticausal pencil as

$$D_m(z) = [E_S - zA_S \quad zB_S]. \quad (3.47)$$

Define also the closed-loop induced pencils of a well-defined induced partition as

$$P_m^l(s) = [sE_T - (A_T + B_T K_T)], \quad (3.48)$$

and

$$D_m^l(z) = [(E_S + B_S K_S) - zA_S]. \quad (3.49)$$

The properties of restricted pencils have been studied extensively (Gantmacher, 1959; Jaffe and Karcanias, 1981; Van Dooren, 1979, 1981). In our approach we point out some of these properties in order to exhibit their significance in the problem of PD feedback.

We have already pointed out that derivative feedback is used on \mathcal{Y} . Equation (3.12) indicates that the finite poles of the closed-loop that are assigned with derivative feedback are given by the reciprocals of the nonzero eigenvalues of F_S . Moreover, the poles at infinity have degrees equal to the lengths, minus 1, of the zero eigenvector chains of F (i.e. the sizes of the zero Jordan blocks of F_S minus one). That is, on \mathcal{Y} the finite closed-loop poles are the reciprocals of the finite modes of $D_m^l(z)$, while the closed-loop modes at infinity are given by the zero elementary divisors, minus 1, of $D_m^l(z)$. This definition is consistent with the point of view of Pugh and Ratcliff (1979).

We may now explore the possibility of selecting the closed-loop modes on a given partition $(\mathcal{Y}, \mathcal{F})$ by PD feedback, which amounts to selecting F_S and F_T with desired eigenvalues in (2.10) and (2.11), or equivalently in (2.28) and (2.29), once $(\mathcal{Y}, \mathcal{F})$ have been specified.

Our technique of solving the generalized Lyapunov equations requires the knowledge of the unreachable modes of the induced pencils (3.46) and (3.47). The matrix pencil theory along with the geometric theory can solve this problem.

Define the *finite unreachable* modes of (3.46) as

$$\sigma_u^l(E_T, A_T) = \{\alpha \in \sigma(E_T, A_T) \mid \text{rank}(P_m(s)) < r\}, \quad (3.50)$$

and the *finite reachable* modes of (3.46) as

$$\sigma_r^l(E_T, A_T) = \{\alpha \in \sigma(E_T, A_T) \mid \text{rank}(P_m(s)) = r\}. \quad (3.51)$$

Similarly, define the *finite unreachable* modes of

(3.47) as

$$\sigma_u^y(E_S, A_S) = \{\alpha \in \sigma(E_S, A_S) \mid \text{rank}(D_m(z)) < \sigma\}, \quad (3.52)$$

and the finite reachable modes of (3.47) as

$$\sigma_r^y(E_S, A_S) = \{\alpha \in \sigma(E_S, A_S) \mid \text{rank}(D_m(z)) = \sigma\}. \quad (3.53)$$

By using the algorithm proposed by Van Dooren (1981) the induced pencils (3.46) and (3.47) $P_m(s)$ and $D_m(z)$ can be transformed to the form

$$\begin{pmatrix} sE_T^r - A_T^r & 0 & 0 & 0 \\ * & sE_T^t - A_T^t & 0 & 0 \\ * & sE_T^1 - A_T^1 & sE_T^0 - A_T^0 & B_0 \end{pmatrix} \quad (3.54)$$

$$\begin{pmatrix} E_S^r - zA_S^r & 0 & 0 & 0 \\ * & E_S^t - zA_S^t & 0 & 0 \\ * & E_S^1 - zA_S^1 & E_S^0 - zA_S^0 & zB_0 \end{pmatrix} \quad (3.55)$$

where * denotes possibly nonzero entries, $[sE_T^t - A_T^t]$, $[E_S^t - zA_S^t]$ and $[sE_T^r - A_T^r]$, $[E_S^r - A_S^r]$ are (square) regular pencils containing the finite and infinite elementary divisors, and the pencils

$$[sE_T^0 - A_T^0 \quad B_0] \quad (3.56)$$

and

$$[E_S^0 - zA_S^0 \quad zB_0] \quad (3.57)$$

are of full rank for all s, z with $[sE_T^r - A_T^r]$ and $[E_S^r - zA_S^r]$ regular pencils. Moreover (3.56) and (3.57) are in the generalized Hessenberg form.

Since the pencils $sE_T^t - A_T^t$ and $E_S^t - zA_S^t$ contain the finite elementary divisors of the induced pencils (3.46) and (3.47) then the spectra of these pencils correspond to the unreachable modes of the pencils (3.46) and (3.47) respectively (Karcianas and Kalogeropoulos, 1987). Moreover the anticausal pencil deals with the reciprocals of these modes. That is, its zero unreachable mode corresponds to the unreachable mode at infinity for the system pencil (3.44). For further details see Karcianas and Kalogeropoulos (1987).

3.3.2. The solution to the uncoupled generalized Lyapunov equations. By restricting (3.54) and (3.55) to \mathcal{T}_r^* and \mathcal{S}_r^* and proposing a lower triangular form for T, S and a diagonal form for F_r, F_s then (3.54) and (3.55) become

$$\begin{pmatrix} A_T^r & 0 \\ A_T^1 & A_T^0 \end{pmatrix} \begin{pmatrix} T_r^r & 0 \\ T_r^1 & T_r^0 \end{pmatrix} - \begin{pmatrix} E_T^r & 0 \\ E_T^1 & E_T^0 \end{pmatrix} \begin{pmatrix} T^r & 0 \\ T^1 & T^0 \end{pmatrix} \begin{pmatrix} F_r^r & 0 \\ 0 & F_r^0 \end{pmatrix} = - \begin{pmatrix} 0 \\ B_T^0 \end{pmatrix} (G_T^r \quad G_T^0) \quad (3.58)$$

and

$$\begin{pmatrix} E_S^r & 0 \\ E_S^1 & E_S^0 \end{pmatrix} \begin{pmatrix} S_o^r & 0 \\ S_o^1 & S_o^0 \end{pmatrix} - \begin{pmatrix} A_S^r & 0 \\ A_S^1 & A_S^0 \end{pmatrix} \begin{pmatrix} S_o^r & 0 \\ S_o^1 & S_o^0 \end{pmatrix} \begin{pmatrix} F_s^r & 0 \\ 0 & F_s^0 \end{pmatrix} = - \begin{pmatrix} 0 \\ B_S^0 \end{pmatrix} (G_S^r \quad G_S^0) \quad (3.59)$$

Observe that (3.58) and (3.59) are the uncoupled generalized Lyapunov equations. In the following theorem we propose a solution to these equations. Moreover, the closed-loop desired modes have been assigned to the closed-loop system restricted to \mathcal{S}, \mathcal{T} . This design technique is based on the solution of these equations. The next theorem is of practical importance since it demonstrates such feedbacks and also guarantees the closed-loop regularity.

Theorem 9. Define

$$p = \dim E_{\mathcal{T}} \mathcal{T}_r^* \quad (3.60a)$$

$$d = \dim A_S \mathcal{S}_r^*. \quad (3.60b)$$

Given a desired closed system structure select a self-conjugate set Σ_p of p desired modes of $P_m^r(s)$ and a self-conjugate set Σ_d of d reciprocal (note that $z = s^{-1}$) desired modes of $D_m^r(z)$ such that

$$\sigma_u^r(E, A) \subset \Sigma_p \quad (3.61a)$$

$$\sigma_u^s(E, A) \subset \Sigma_d. \quad (3.61b)$$

Suppose moreover that

$$\sigma_u^r(E, A) \cap \sigma_r^r(E, A) = 0 \quad (3.62a)$$

$$\sigma_u^s(E, A) \cap \sigma_r^s(E, A) = 0 \quad (3.62b)$$

that is no finite reachable mode under PD feedback has the same value as a finite unreachable mode under PD feedback.

Then there exists a feedback (3.1) that assigns Σ_p and Σ_d as the closed-loop modes of $P_m^r(s)$ and $D_m^r(z)$ and ensures the regularity of (3.2) on a partition \mathcal{S}, \mathcal{T} , such that

$$E_{\mathcal{T}} \mathcal{T}_r \subset E_{\mathcal{T}} \mathcal{T}_r^* \quad (3.63a)$$

$$A_S \mathcal{S}_r \subset A_S \mathcal{S}_r^*. \quad (3.63b)$$

Proof. This proof is a combination of the proofs given in Lewis and Özçaldıran (1989) and Lewis and Syrmos (1991) for pure proportional and pure derivative feedback. We have to emphasize that the proof is constructive and based on the solution of (3.58) and (3.59). We also note that this exploits the generalized Hessenberg form of (3.56) and (3.57). The computational technique is stable since it involves only unitary transformations as defined by Verhagen and Van Dooren (1986). The regularity of the pencils

$P_m^d(s)$ and $D_m^d(z)$ is guaranteed in Lewis and Özcaldiran (1989) and Lewis and Syrmos (1991). Since, now this is true, Theorem 7 guarantees the closed-loop regularity of (3.2). This is true since we work with a well-defined induced partition. Hence at this point the reason of the selection of such a partition is obvious. This argument justifies the specific selection which was not evident in the beginning of this section.

The computation of p and d in Theorem 9 reveals the maximum number of poles that can be assigned by PD feedback on $\mathcal{F} = \mathcal{Y} \odot \mathcal{T}$. We claim now that the finite closed-loop spectrum that is assigned to the pencils $[s(E_S + B_S K_S) - A_S]$ and $[sE_T - (A_T + B_T K_T)]$, that is $P_m^d(s)$ and $sD_m^d(s^{-1})$ is the same one that is assigned to the pencil $[s(E + BK_2) - (A + BK_1)]$. Actually the following Lemma shows that this is indeed the case.

Lemma 7. Let $\mathcal{F} = \mathcal{Y} \odot \mathcal{T}$ be a regular partition and $K_1 S = K_2 T = 0$. Then the finite spectrum of the pencil $[s(E + BK_2) - (A + BK_1)]$ is the union of the spectra of the pencils $[s(E_S + B_S K_S) - A_S]$ and $[sE_T - (A_T + B_T K_T)]$ restricted to \mathcal{Y} and \mathcal{T} respectively, where E_S , A_S , E_T , A_T are the matrices that are constructed by Algorithm 1, that is they constitute a well-defined induced partition.

Proof. Consider the projection matrix Q as follows

$$Q = \begin{pmatrix} Q_S \\ Q_T \end{pmatrix} \quad (3.64)$$

where $Q \in \mathcal{R}^{n \times n}$, $Q_S \in \mathcal{R}^{\sigma \times n}$ and $Q_T \in \mathcal{R}^{\tau \times n}$. Then Q_S and Q_T , since they are canonical projections as defined in Theorem 7, are of full row rank matrices. Specifically $\text{rank } Q_S = \sigma$ and $\text{rank } Q_T = \tau$. Then we can write

$$\begin{aligned} Q[s(E + BK_2) - (A + BK_1)][s \quad T] \\ &= Q[s(E + BK_2) - (A + BK_1)S \quad s(E + BK_2)T \\ &\quad - (A + BK_1)T] \\ &= Q[s(E + BK_2) - AS \quad sET - (A + BK_1)T] \\ &= \begin{pmatrix} Q_S \\ Q_T \end{pmatrix} (s(E + BK_2)S - AS \quad sET \\ &\quad - (A + BK_1)T). \end{aligned} \quad (3.65)$$

By premultiplying (3.3), (3.4) by Q_S , Q_T respectively and by using (2.18–2.1) we get

$$(A_S P_S + B_S K_1) \mathcal{F} \subset E_S P_S \mathcal{F} \quad (3.66)$$

$$(E_T P_T + B_T K_2) \mathcal{F} \subset A_T P_T \mathcal{F}. \quad (3.67)$$

But by the definition of P_S , P_T we know that

$P_S T = P_T S = 0$, therefore (3.66) and (3.67) become as follows

$$B_S K_1 \mathcal{F} \subset 0 \quad (3.68)$$

$$B_T K_2 \mathcal{F} \subset 0, \quad (3.69)$$

or

$$B_S K_1 T = 0 \quad (3.70)$$

$$B_T K_2 S = 0. \quad (3.71)$$

Direct manipulation of (3.65) using (2.18–21), (2.30), (2.31), (3.70), and (3.71) yields

$$\begin{pmatrix} [s(E_S + B_S K_S) - A_S] S, & 0 \\ 0 & [sE_T - (A_T + B_T K_T)] T_r \end{pmatrix} \\ \times \begin{pmatrix} S_o & 0 \\ 0 & T_r \end{pmatrix}. \quad (3.72)$$

Then by using (3.38) and (3.39), (3.72) can be written as follows

$$\begin{pmatrix} [s(E_S + B_S K_S) - A_S] & 0 \\ 0 & sE_T - (A_T + B_T K_T) \end{pmatrix} \\ \times P[S \quad T] \quad (3.73)$$

Therefore these pencils are related as follows

$$\begin{aligned} Q[s(E + BK_2) - (A + BK_1)] \\ &= \begin{pmatrix} s(E_S + B_S K_S) - A_S & 0 \\ 0 & sE_T - (A_T + B_T K_T) \end{pmatrix} P. \end{aligned} \quad (3.74)$$

But by the construction of the well-defined induced partition we know that $\text{rank } P = \text{rank } Q = \sigma + \tau$. Therefore these pencils have the same finite spectrum. ■

At this point, it is evident that solving the uncoupled Lyapunov equations is equivalent to solving the coupled Lyapunov equations. Moreover the pole assignment using a well-defined induced partition and the induced closed-loop pencils has been considered.

4. FEEDBACK-FREE INDUCED PARTITIONS AND SPECTRUM ASSIGNABILITY

The approach using matrix pencil equivalence ideas and quotient subspaces concepts challenges us to consider the closed-loop spectrum assignability by using feedback-free induced mapping partitions. In this section we study the structure of feedback-free description of (2.1), that is (2.32). Based on this information we present methods for spectrum assignability. Note that these methods are "feedback-free" in the sense that they do not depend on B . This method involves reduced order nonsquare pencils. These

pencils can be easily handled and studied. Specifically, the theory of singular systems covers the theory of nonsquare pencils. The solution to the spectrum assignability problem is based on the feedback-free uncoupled generalized Lyapunov equations. These equations follow from the feedback-free induced pencils. As a result they involve nonsquare reduced order matrices. Therefore the computational stability of the algorithm is better than the one proposed in Section 3. Moreover, since these equations are independent of B , alternative computational techniques can be used.

4.1. Feedback-free partitions and the closed-loop system

For our purposes we will need some ancillary results that exhibit our motivation and constitute the cornerstone for further exploration. In these results we do not assume the property of open-loop regularity. The next theorem shows how the restricted pencil to a partition (S, T) can be decomposed into a form that involves the concept of the feedback-free induced partition.

Theorem 10. Let (S, T) be a partition. Then there always exists a feedback-free induced partition such that

$$\tilde{Q}(sL - M) = \begin{pmatrix} sL_S - M_S & 0 \\ 0 & sL_T - M_T \end{pmatrix} P. \quad (4.1)$$

Proof. Consider

$$\tilde{Q} = \begin{pmatrix} Q_S \\ \tilde{Q}_T \end{pmatrix} \quad (4.2)$$

to be the projection map where \tilde{Q}_T and \tilde{Q}_S are defined in Lemmas 5 and 6. Let now N be the left annihilator of B as defined in Section 2. Then by using (2.38–36) the following hold true

$$\begin{pmatrix} \tilde{Q}_S \\ \tilde{Q}_T \end{pmatrix} (sL - M) = \begin{pmatrix} sL_S P_S - M_S P_S \\ sL_T P_T - M_T P_T \end{pmatrix}$$

or

$$\begin{aligned} \tilde{Q}(sL - M) &= \begin{pmatrix} sL_S - M_S & 0 \\ 0 & sL_T - M_T \end{pmatrix} \begin{pmatrix} P_S \\ P_T \end{pmatrix} \\ &= \begin{pmatrix} sL_S - M_S & 0 \\ 0 & sL_T - M_T \end{pmatrix} P. \end{aligned} \quad (4.3)$$

Thus (4.1) holds true for every feedback-free partition. ■

For our purposes, that is spectrum assignability, we require regularity for the closed-loop system. This concept leads us to focus on the class of regular partitions. This effect is exhibited in the next theorem.

Theorem 11. Define a regular partition (S, T) such that $\dim \mathcal{T} = \dim \tilde{\mathcal{T}}$ and $\dim \mathcal{S} = \dim \tilde{\mathcal{S}}$. Then there exists a feedback-free partition such that

$$\begin{aligned} \tilde{Q}(sL - M) &= \begin{pmatrix} N_S & 0 \\ 0 & N_T \end{pmatrix} \\ &\times Q[s(E + BK_2) - (A + BK_1)], \end{aligned} \quad (4.4)$$

where $[s(E + BK_2) - (A + BK_1)]$ is regular.

Proof. Since (S, T) is a regular partition, then there exist K_1, K_2 that satisfy (3.8), (3.9) such that the closed-loop system is regular. Moreover we can always construct a well-defined partition such that (3.74) holds. By premultiplying (3.74) by

$$\begin{pmatrix} N_S & 0 \\ 0 & N_T \end{pmatrix} \quad (4.5)$$

we get

$$\begin{aligned} \begin{pmatrix} N_S & 0 \\ 0 & N_T \end{pmatrix} Q[s(E + BK_2) - (A + BK_1)] \\ = \begin{pmatrix} sL_T - M_T & 0 \\ 0 & sL_S - M_S \end{pmatrix} P. \end{aligned} \quad (4.6)$$

Then by using (4.1) and (4.6) we observe that (4.3) holds true where $[s(E + BK_2) - (A + BK_1)]$ is regular. ■

The next theorem reveals another property for the regularity of the closed-loop system by providing conditions for the choice of \tilde{S} and \tilde{T} . This selection will later give us the ability to transform this feedback-free description form to a regular closed-loop form description.

Theorem 12. Define a partition $\mathcal{X} = \mathcal{S} \odot \mathcal{T}$ such that \mathcal{T} and \mathcal{S} are (L, M) -deflating subspaces. Then the pencil $sL - M$ restricted to \mathcal{X} is of full row rank if and only if the pencils $sL_S - M_S$ and $sL_T - M_T$ restricted to \mathcal{S} and \mathcal{T} are of full row rank, where L_T, M_T, L_S , and M_S define a feedback-free induced partition.

Proof. Note that since (2.37) and (2.38) hold, then the following is true

$$\begin{pmatrix} N_S & 0 \\ 0 & N_T \end{pmatrix} \begin{pmatrix} Q_S \\ Q_T \end{pmatrix} = \begin{pmatrix} N_S Q_S \\ N_T Q_T \end{pmatrix} = \begin{pmatrix} \tilde{Q}_S \\ \tilde{Q}_T \end{pmatrix} N. \quad (4.7)$$

Choose as in Theorem 5 $\{t_i\}_{i=1}^r, \{s_i\}_{i=1}^o, \{x_i\}_{i=1}^x$ be bases for \mathcal{T}, \mathcal{S} and \mathcal{X} respectively. Then by using (4.1), (4.4) and (4.7) we get

$$(sL_S - M_S)s_{\sigma} = N_S Q_S w_{1+\sigma}, \quad i \in \{1, \dots, \sigma\} \quad (4.8)$$

$$(sL_T - M_T)t_{\tau} = N_T Q_T w_i, \quad i \in \{1, \dots, \tau\} \quad (4.9)$$

where clearly $\{s_{\alpha}\}_{\alpha=1}^{\sigma}$ and $\{t_{\tau}\}_{\tau=1}^{\tau}$ are bases for \mathcal{S} and \mathcal{T} respectively. Then using (2.37) and (2.38) we get

$$(sL_S - M_S)s_{\alpha} = \tilde{Q}_S N w_{1+\alpha}, \quad i \in \{1, \dots, \sigma\} \quad (4.10)$$

$$(sL_T - M_T)t_{\tau} = \tilde{Q}_T N w_{1+\tau}, \quad i \in \{1, \dots, \tau\}. \quad (4.11)$$

It follows from (4.10) and (4.11) that the pencils $sL_S - M_S$ and $sL_T - M_T$ have full row rank restricted to \mathcal{S} and \mathcal{T} if and only if $\dim \mathcal{S} = \dim \tilde{\mathcal{S}}$ and $\dim \mathcal{T} = \dim \tilde{\mathcal{T}}$, which is always true by assumption. The converse follows easily as in Theorem 7. ■

Theorem 7 in the previous section indicates how to construct a well-defined induced partition. Similarly, Theorem 12 indicates how to construct a *well-defined feedback-free* induced partition. The condition that requires the partition to be regular will reveal its importance at the end of our design technique while we try to build back our closed-loop system description (3.2). Once having computed a feedback-free well-defined partition, the next step will be how to find a corresponding well-defined induced partition. This will later lead us to a regular closed-loop system with the desired closed-loop behavior. The next theorem provides the relationship between a well-defined feedback free partition and a well-defined induced partition.

Theorem 13. Let $\mathcal{X} = \mathcal{S} \oplus \mathcal{T}$ be a partition. Then given a well-defined feedback-free partition constructed as in Theorem 12 there always exists a well-defined induced partition as defined in Theorem 7 so that

$$\begin{aligned} & \begin{pmatrix} sL_S - M_S & 0 \\ 0 & sL_T - M_T \end{pmatrix} P \\ &= \begin{pmatrix} N_S & 0 \\ 0 & N_T \end{pmatrix} \\ & \quad \times \begin{pmatrix} s(E_S + B_S K_S) - A_S & 0 \\ 0 & sE_T - (A_T + B_T K_T) \end{pmatrix} \\ & \quad \times P. \end{aligned} \quad (4.12)$$

Proof. Direct manipulation of the RHS of (4.12) gives

$$\begin{aligned} & \begin{pmatrix} sL_S - M_S & 0 \\ 0 & sL_T - M_T \end{pmatrix} P \\ &= \begin{pmatrix} \tilde{Q}_S \\ \tilde{Q}_T \end{pmatrix} (sL - M) \\ &= \begin{pmatrix} \tilde{Q}_S \\ \tilde{Q}_T \end{pmatrix} N [s(E + BK_2) - (A + BK_1)] \end{aligned}$$

$$\begin{aligned} &= \begin{pmatrix} N_S & 0 \\ 0 & N_T \end{pmatrix} Q [s(E + BK_2) - (A + BK_1)] \\ &= \begin{pmatrix} N_S & 0 \\ 0 & N_T \end{pmatrix} \\ & \quad \times \begin{pmatrix} s(E_S + B_S K_S) - A_S & 0 \\ 0 & sE_T - (A_T + B_T K_T) \end{pmatrix} \\ & \quad \times P. \end{aligned} \quad (4.13)$$

Note first that K_1 and K_2 have to satisfy (3.7) and (3.8), but this will not attract our concern now since it is an intermediate step at this point. We will later discuss the choice of K_1 and K_2 . Our concern is to verify that the maps A_S, E_S, A_T, E_T constitute a well defined partition. Note that the selection of the well-defined feedback-free partition in Theorem 12 guarantees that the maps A_S, E_S, A_T, E_T constitute a well-defined partition. This follows also by applying rank criteria for Q_S and Q_T using (2.37) and (2.38) and exploiting the fact that this is a well-defined feedback-free partition constructed as in Theorem 12. ■

In light of these concepts the classes of feedback-free partitions depend on the selection of $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$. Among these classes select the one which is addressed by the proof of Theorem 7.

Once we construct the desired feedback-free induced partition using Lemma 4 and taking under consideration the conditions of Theorem 12, we are ready to proceed with our technique. After that we have to be sure that we can find our way back to the closed-loop system without destroying the closed-loop regularity. Theorem 13 guarantees this procedure.

We have exhibited how we can go from a partition to a well-defined feedback-free induced partition. Then Theorem 13 shows how to go from a well-defined feedback-free induced partition to a well-defined induced partition which in the sequel will lead us to the desired closed-loop system. Since we ensure this procedure we are ready to show how a well-defined feedback-free induced partition can be used for the spectrum assignability problem. In this attempt we need some preliminary results as in Section 3.

4.2. The spectrum assignability problem using nonsquare matrices

The subspaces \mathcal{S}, \mathcal{T} of a well-defined feedback-free partition are (L_S, M_S) - and (M_T, L_T) -invariant subspaces and are closed under addition, so that each one has a largest member. We symbolize the *supremal* (L_S, M_S) -invariant subspace as \mathcal{S}_α^* and the *supremal* (M_T, L_T) -invariant subspace as \mathcal{T}_τ^* . The next theorem shows how to compute \mathcal{S}_α^* and \mathcal{T}_τ^* .

Theorem 14. Consider the following subspace recursions

$$\mathcal{X}_{k+1} = M_T^{-1}(L_T \mathcal{X}_k), \quad \text{with } \mathcal{X}_0 = \mathcal{R}^* \quad (4.14)$$

$$\mathcal{Y}_{k+1} = L_S^{-1}(M_S \mathcal{Y}_k), \quad \text{with } \mathcal{Y}_0 = \mathcal{R}^o \quad (4.15)$$

then $\mathcal{F}_0^* = Y_0$ and $\mathcal{F}_1^* = X_1$. ■

As an alternative to (4.14) and (4.15) \mathcal{F}_0^* and \mathcal{F}_1^* may be computed by the numerically convenient Singular System Structure Algorithm (Lewis, 1986).

Consider a regular partition $(\mathcal{F}, \mathcal{T})$. Define the *causal pencil* of (2.32) as

$$P(s) = [sL - M], \quad (4.16)$$

and the *anticausal pencil* of (2.32) as

$$D(z) = [L - zM], \quad (4.17)$$

where $z = s^{-1}$. Having computed as in Theorem 12 a well-defined regular partition, then define the feedback-free induced causal pencil as

$$P_n(s) = [sL_T - M_T], \quad (4.18)$$

and the anticausal pencil as

$$D_n(z) = [L_S - zM_S]. \quad (4.19)$$

We should point out that the unreachable modes of the pencils (3.44) and (3.45) are invariant under the action of N (Karacanias and Kalogeropoulos, 1987). As a result of this remark we can state that the unreachable modes (which correspond to the finite elementary divisors of a pencil) of (3.46) and (3.47) are the same as those of the pencils (4.18) and (4.19). Despite this, we will symbolize them as $\sigma_u^T(L, M)$ and $\sigma_u^o(L, M)$ respectively for compatibility with our notation of the L, M matrices though they are nonsquare. Using the same computational technique for the transformation of the pencils (4.18) and (4.19) to their Kronecker canonical form, and restricting them to \mathcal{T}_1^* and \mathcal{T}_0^* respectively we get the following set of generalized Lyapunov equations

$$\begin{pmatrix} M_T^T & 0 \\ M_T^1 & M_T^0 \end{pmatrix} \begin{pmatrix} T_1^T & 0 \\ T_1^1 & T_1^0 \end{pmatrix} = \begin{pmatrix} L_T^T & 0 \\ L_T^1 & L_T^0 \end{pmatrix} \begin{pmatrix} T_1^T & 0 \\ T_1^1 & T_1^0 \end{pmatrix} \begin{pmatrix} F_T^T & 0 \\ 0 & F_T^0 \end{pmatrix} \quad (4.20)$$

and

$$\begin{pmatrix} L_S^T & 0 \\ L_S^1 & L_S^0 \end{pmatrix} \begin{pmatrix} S_0^T & 0 \\ S_0^1 & S_0^0 \end{pmatrix} = \begin{pmatrix} M_S^T & 0 \\ M_S^1 & M_S^0 \end{pmatrix} \begin{pmatrix} S_0^T & 0 \\ S_0^1 & S_0^0 \end{pmatrix} \begin{pmatrix} F_S^T & 0 \\ 0 & F_S^0 \end{pmatrix}. \quad (4.21)$$

Observe that (4.20) and (4.21) are reduced order uncoupled generalized Lyapunov equations. Therefore we significantly reduce the order of

the original matrices. It is evident now that the solution of these equations is much easier than those of (3.58) and (3.59) due not only to the fact that are reduced order equations but also to the fact that they are invariant of the presence of B . Thus we reduced the amount of required operations. Also, the algorithm becomes less complicated and easier to implement. We note that the B has not been annihilated. On the contrary, it plays a significant role through the following equations.

$$B_T^{0*}(A_T^{-1} A_T^0)(T_1^T \ T_1^0) - B_T^{0*}(E_T^1 \ E_T^0) \times (T_1^1 \ T_1^0) \begin{pmatrix} 0 & F_T^0 \end{pmatrix} = -(G_T^T \ G_T^0) \quad (4.22)$$

and

$$B_S^{0*}(E_S^1 \ E_S^0)(S_0^1 \ S_0^0) - B_S^{0*}(A_S^1 \ A_S^0) \times (S_0^1 \ S_0^0) \begin{pmatrix} 0 & F_S^0 \end{pmatrix} = -(G_S^T \ G_S^0) \quad (4.23)$$

where B_T^{0*} and B_S^{0*} are the Moore–Penrose inverses of B_T^0 and B_S^0 .

We point out that these equations consist only of matrix–matrix multiplication operations. Therefore given the F_S and F_T the solution of the feedback-free generalized Lyapunov equations, that is S and T completely defines G_S and G_T through the equations (4.22) and (4.23). Under these considerations our goal is to find \mathcal{F} and \mathcal{T} using the computationally easier equations (4.20) and (4.21), then substituting \mathcal{F} and \mathcal{T} back to (4.22) and (4.23) we compute G_S and G_T . Then using G_S and G_T as well as our ancillary results of this section we will define the feedback gains K_1 and K_2 that guarantee the closed-loop regularity of (3.2) and also assign the desired spectrum. For these reasons the next theorem explains how we solve (4.20) and (4.21) for the specific selection of F_S and F_T . This theorem is the corresponding result of that of Theorem 9.

Theorem 15. Define

$$p \equiv \dim L_T \mathcal{T}_1^* \quad (4.24a)$$

$$d \equiv \dim M_S \mathcal{T}_0^*. \quad (4.24b)$$

Given a desired closed system structure select a self-conjugate set Σ_p of p desired modes of $P_{in}^T(s)$ and a self-conjugate set Σ_d of d reciprocal (note that $z = s^{-1}$) desired modes of $D_{in}^T(z)$ such that

$$\sigma_u^T(L, M) \subset \Sigma_p \quad (4.25a)$$

$$\sigma_u^o(L, M) \subset \Sigma_d. \quad (4.25b)$$

Suppose moreover that

$$\sigma_u^T(L, M) \cap \sigma_r^T(L, M) = 0 \quad (4.26a)$$

$$\sigma_u^o(L, M) \cap \sigma_r^o(L, M) = 0 \quad (4.26b)$$

that is no finite reachable mode under PD feedback has the same value as a finite unreachable mode under PD feedback.

Then there exists a feedback (3.1) that assigns Σ_p and Σ_d as the closed-loop modes of $P_m'(s)$ and $D_m^d(z)$ and ensures the regularity of (3.2) on a partition \mathcal{S}, \mathcal{T} , such that

$$L_T \mathcal{T}_t \subset L_T \mathcal{T}_t^* \quad (4.27a)$$

$$M_S \mathcal{S}_o \subset M_S \mathcal{S}_o^*. \quad (4.27b)$$

Proof. The first part of the proof is based on the solution of the uncoupled feedback-free generalized Lyapunov equations. This solution is a special case ($B=0$) of the one proposed in Theorem 9. It also exploits all the advantages by using unitary transformations. Moreover, it is based on the generalized Hessenberg form (Verhagen and Van Dooren, 1986). There are several advantages of this technique. The order of the matrices is reduced, the algorithm is less complicated and can therefore be more easily implemented, and finally G_S and G_T are computed explicitly using the equations (4.22) and (4.23).

The second part of the proof is based on the construction of feedback gains that uses the solutions of these Lyapunov equations while the closed-loop system is regular. Moreover we will verify how these feedback gains assign the desired closed-loop spectrum.

Since we used a well-defined feedback-free partition, then according to Theorem 13 we know that (4.17) holds true. Therefore we need to calculate the projections Q_S and Q_T in order to compute the well-defined induced partition that is defined by the induced maps E_S, A_S, E_T, A_T . This can be easily accomplished by using (2.37) and (2.38). Having computed Q_S and Q_T we calculate all the induced maps E_S, A_S, E_T, A_T as defined in Theorem 1. This is a well-defined induced mapping partition and is guaranteed by Theorem 13.

Our next step is to construct feedback gains from the computed G_T, G_S and T_t, S_o . This can be easily done using the following

$$G_T = K_T T_t \quad (4.28)$$

$$G_S = K_S S_o. \quad (4.29)$$

Then we can also compute K_1 and K_2 as follows

$$K_1 = K_T P_T \quad (4.30)$$

$$K_2 = K_S P_S. \quad (4.31)$$

But these feedback gains assign the desired closed-loop spectrum and guarantee the closed-loop regularity as Theorem 7 shows. ■

In order to summarize this technique we provide the most critical steps in the following algorithm.

Algorithm 2

Step 1: Define a partition $(\mathcal{S}, \mathcal{T})$ such that (3.26), (3.27) hold true.

Step 2: Perform row operations on B as defined in (3.28–31) and compute $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$ using (3.18) and (3.19).

Step 3: Compute now a well-defined feedback-free induced partition using the $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$ as defined in Theorem 13.

Step 4: Solve the feedback-free generalized Lyapunov equations for the given closed-loop specifications.

Step 5: For the computed $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$ calculate P_S, Q_S and P_T, Q_T as defined in Theorem 1.

Step 6: Using equations (2.18–21) compute E_S, A_S, E_T, A_T , that is a well-defined induced mapping partition.

Step 7: Using equation (4.28–31) compute the necessary gains for the desired closed-loop behavior. ■

5. AN EXAMPLE

We consider the system defined by

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} -2 & 0 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} x + \begin{pmatrix} 6 \\ 3 \\ -2 \end{pmatrix} u. \quad (5.1)$$

This singular system has been derived by Newcomb and Dziourla (1989) for the problem of "input for specific output". In order to avoid lengthy calculations we first bring (5.1) to the generalized lower Hessenberg form by using the algorithm proposed by Van Dooren (1979). This yields

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -2 & -2 \\ 0 & 0 & 1 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix} u. \quad (5.2)$$

We will use $\mathbf{R}(\cdot)$ and e_i to denote the range of a subspace and the i th column of the identity matrix.

Let $\mathcal{S} = \mathbf{R}(e_1)$ and $\mathcal{T} = \mathbf{R}(e_2, e_3)$. It is easy to check that $\mathcal{K} = \mathcal{S} \odot \mathcal{T}$ defines a partition. According now to Theorem (7) note that $\tilde{\mathcal{T}} = \mathbf{R}(e_2, e_3)$ and $\tilde{\mathcal{S}} = \mathbf{R}(e_1)$ define a well-defined induced mapping partition $\mathcal{K} = \mathcal{S} \odot \mathcal{T}$ such that $\dim \tilde{\mathcal{T}} = \dim \tilde{\mathcal{S}} = 2$ and $\dim \mathcal{S} = \dim \tilde{\mathcal{S}} = 1$ for $H_T = 0$ and $H_S = 0$, where $\tilde{\mathcal{T}} = \mathbf{R}(e_2, e_3)$ and $\tilde{\mathcal{S}} = \mathbf{R}(e_1)$. Using equations (2.18)–(2.21) the induced maps are

$$E_S = 0, \quad \tilde{E}_T = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

$$A_S = -1, \quad A_T = \begin{pmatrix} -2 & -2 \\ 0 & 1 \end{pmatrix},$$

$$B_S = 0, \quad B_T = \begin{pmatrix} 0 \\ -2 \end{pmatrix}.$$

Moreover the induced pencils are

$$D_m(s) = [-z \quad 0]$$

$$P_m(z) = \begin{pmatrix} s+2 & 2 & 0 \\ 0 & -1 & -2 \end{pmatrix}.$$

It is easy to see that $[sE_T - A_T \quad B_T]$ contains one finite reachable mode, actually the mode $\{-2\}$. The pencil $[E_S - zA_S \quad B_S]$ contains one finite unreachable mode, actually the mode $\{0\}$. It is noteworthy to mention that the original system restricted to \mathcal{T} and \mathcal{F} respectively contains one column minimal index of order one and one infinite elementary divisor (ied) respectively. The ied of the original pencil has been transformed to a zero elementary divisor as was expected.

In order now to solve the uncoupled Lyapunov equations we compute $d = \dim A_S \mathcal{F}_\alpha^* = 0$ and $p = \dim E_T \mathcal{T}_\tau^* = 1$. Let now $\{-4\}$ be the desired spectrum. Since the pencil $P_m(s)$ is already in its Hessenberg form, then equation (3.58) can be written as

$$\begin{pmatrix} 2t_1 - 2t_2 \\ t_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} g_r,$$

a solution to which is $t_1 = t_2 = 1$ and $g_r = \frac{1}{2}$. Therefore by using $g_r = K_r T_r$ we find that $K_r = [\frac{1}{2} \quad 0]$. It is easy to check that the spectrum of the subsystem is $\{-4\}$. Moreover, $K_1 = K_r P_r = [0 \quad \frac{1}{2} \quad 0]$.

Similarly, by solving (3.59) for the pencil $D_m(s)$, we get $S_o = 1$ and G_s can be chosen arbitrarily. Therefore select as $K_s = 1$, then $K_2 = [1 \quad 0 \quad 0]$. We can see that the spectrum of the closed-loop subsystem is $\{0\}$. Knowing that the feedback gains are $K_2 = [1 \quad 0 \quad 0]$ and $K_1 = [0 \quad \frac{1}{2} \quad 0]$, we can verify that the finite closed-loop spectrum is $\{-4\}$. It is also noteworthy to observe that the Kronecker structure of the system remained invariant under the proposed decomposition.

The same problem can be alternately confronted by using the feedback-free approach. We will also use in this case the transformed system (5.2). In order to do so select

$$N = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad B^+ = (0 \quad 0 \quad -\frac{1}{2})$$

such that $NB = 0$ and $B^+ = 1$. Then

$$L = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad M = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -2 & -2 \end{pmatrix}.$$

By choosing again the same partition as previously employed to be a feedback-free induced partition, the induced maps are

$$L_S = 0, \quad L_T = (1 \quad 0),$$

$$M_S = -1, \quad M_T = (-2 \quad -2).$$

Consequently the induced pencils are

$$D_m(z) = [-z]$$

$$P_m(s) = (s+2 \quad 2).$$

In order to emphasize the importance of the discussion on p. 366, we point out that the pencil $D_m(z)$ contains a zero finite elementary divisor, while the pencil $P_m(s)$ contains a column minimal index of length two.

In order to solve the uncoupled Lyapunov equations we compute $d = \dim M_S \mathcal{F}_\alpha^* = 0$ and $p = \dim L_T \mathcal{T}_\tau^* = 1$. Let now $\{-4\}$ be the desired spectrum. Since the pencil $P_m(s)$ is already in its Hessenberg form, equation (4.2) can be written as

$$\begin{pmatrix} 2t_1 - 2t_2 \\ t_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} g_r,$$

a solution to which is $t_1 = t_2 = 1$. Moreover, from (4.22) we have that $g_r = \frac{1}{2}$. Therefore by using $g_r = K_r T_r$ we find that $K_r = [\frac{1}{2} \quad 0]$. It is easy to now check that the spectrum of the subsystem is $\{-4\}$. Moreover, $K_1 = K_r P_r = [0 \quad \frac{1}{2} \quad 0]$.

Similarly, by solving (4.21) for the pencil $D_m(s)$, we get $S_o = 1$ and G_s can be chosen arbitrarily. Therefore select as $K_s = 1$, then $K_2 = [1 \quad 0 \quad 0]$. We can see that the spectrum of the closed-loop subsystem is $\{0\}$. Knowing that the feedback gains are $K_2 = [1 \quad 0 \quad 0]$ and $K_1 = [0 \quad \frac{1}{2} \quad 0]$, we can verify that the finite closed-loop spectrum is $\{-4\}$. It is also noteworthy to observe that the Kronecker structure of the system remained invariant under the proposed decomposition.

6. CONCLUSIONS

In this paper we presented a new classification of invariant subspaces by utilizing the concepts of quotient and partitioned subspaces. It was also exhibited how these are related. The partitioned and quotient subspaces preserve certain open-loop properties as was shown in Section 2. Moreover an alternative formulation was used, that is the feedback-free description. This latter one involves the concept of nonsquare pencils.

These properties along with specific conditions that should hold led us later to the determination of a well-defined induced partition. This class of subspaces that is constructed by using the notions of quotient subspaces constituted the

center of interest in Section 3. In this it was presented that this class of subspaces establish a new technique for spectrum assignability using proportional-plus-derivative feedback. The proposed method was based on the decomposition of the closed-loop system into two equivalent subsystems. The solution to the spectrum assignability problem was confronted by utilizing the uncoupled generalized Lyapunov equations. The solution to these equations is based on the generalized Hessenberg form.

Finally in Section 4 the problem of spectrum assignability was confronted from a different point of view. Specifically the concepts of feedback-free induced partition were used. As a result we dealt with nonsquare pencils in order to assign the poles. This technique involved reduced ordered matrices, subsequently the algorithm was computationally stabler. It is noteworthy that the solution of the feedback-free uncoupled generalized Lyapunov equations were invariant of B . Therefore the proposed algorithm was less complicated. The use of nonsquare pencils to the spectrum assignability problem constitutes a powerful technique not only in the singular systems but also in the state-variable systems. At the end the advantages of this method were discussed and an algorithm was given which exhibited the philosophy behind this endeavour.

Acknowledgement—This research was supported by NSF Grant ECS-8805932 and the Alexander Onassis Foundation Group I-89.

REFERENCES

- Basile, G. and G. Marro (1969). Controlled and conditioned invariant subspaces in linear system theory *JOTA*, **4**, 304–315.
- Bernard, P. (1982). On singular implicit linear dynamical systems. *SIAM J. Control Optimiz.* **20**, 612–633.
- Cobb, J. D. (1984). Controllability, observability, and duality in singular systems. *IEEE Trans. Aut. Control*, **AC-29**, 1076–1082.
- Gantmacher, F. R. (1959). *The Theory of Matrices*. New York, Chelsea.
- Jaffe, S. and N. Karcanias (1981). Matrix pencil characterizations of almost (A,B)-invariant subspaces: a classification of geometric concepts. *Int. J. Control*, **33**, 51–93.
- Karcanias, N. and G. Kalogeropoulos (1987). A matrix pencil approach to the study of singular systems: algebraic and geometric aspects. *Proc. Int. Symp. Singular Systems*. Atlanta, GA pp. 29–33.
- Lewis, F. L. (1984). Descriptor systems: decomposition into forward and backward subsystems. *IEEE Trans. Aut. Control*, **AC-29**, 167–170.
- Lewis, F. L. (1986). A survey of linear singular systems. *J. Circuits Syst. Signal Process.* **5**, 3–36.
- Lewis, F. L. and K. Özcaldiran (1989). Geometric structure and feedback in singular systems. *IEEE Trans. Aut. Control*, **AC-34**, 450–455.
- Lewis, F. L. and V. L. Syrmos (1991). A geometric theory for derivative feedback. *IEEE Trans. Aut. Control*, (to appear.)
- Lewis, F. L., M. A. Christodoulou, B. G. Mertzios and K. Özcaldiran (1989). Chained aggregation for singular systems. *IEEE Trans. Aut. Control*, **AC-34**, 1007–1012.
- Malabre, M. (1987). More geometry about singular systems. *Proc. IEEE Conf. on Decision and Control*. Los Angeles, CA, pp. 1138–1139.
- Newcomb, R. W. and B. Dziurla (1989). Some circuits and systems applications of semistate theory. *Circuits Syst. J. Signal Process.* **8**, 235–260.
- Özcaldiran, K. (1985). Control of descriptor systems. Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, GA.
- Özcaldiran, K. (1986). A geometric characterization of the reachable and controllable subspaces of descriptor systems. *J. Circuits Syst. Signal Process.*, Vol. 5, No. 1, pp. 37–48.
- Özcaldiran, K. and F. L. Lewis (1991). On the regularizability of singular systems. *IEEE Trans. Aut. Control* (to appear.)
- Pugh, A. C. and P. A. Ratcliff (1979). On the zeros and poles of a rational matrix. *Int. J. Control*, **30**, 213–226.
- Shayman, M. A. and Z. Zhou (1987). Feedback control characterization and classification of generalized singular systems. *IEEE Trans. Aut. Control*, **AC-32**, 483–496.
- Syrmos, V. L. and F. L. Lewis (1991). Decomposability and quotient subspaces for the pencil $sL - M$. *SIAM J. Matrix Anal. Applic.* (submitted).
- Van Dooren, P. (1979). The computation of Kronecker's canonical form of a singular system. *Lin. Algebra Applic.*, **27**, 103–140.
- Van Dooren, P. (1981). The generalized eigenstructure problem in linear system theory. *IEEE Trans. Aut. Control*, **AC-26**, 111–129.
- Verghese, G. C., B. C. Lévy, and T. Kailath (1981). A generalized state-space for singular systems. *IEEE Trans. Aut. Control*, **AC-26**, 811–831.
- Verghese, G. C., P. Van Dooren, and T. Kailath (1979). Properties of the system matrix of a generalized state-space system. *Int. J. Control*, **30**, 235–243.
- Verhagen, M. H. and P. Van Dooren (1986). A reduced order observer for descriptor systems. *Syst. Control Lett.* **8**, 29–37.
- Willems, J. C. (1981). Almost invariant subspaces: an approach to high gain feedback. Part 1: almost controlled invariant subspaces. *IEEE Trans. Aut. Control*, **AC-26**, 235–252.
- Willems, J. C. (1982). Almost invariant subspaces: an approach to high gain feedback. Part 2: almost conditionally invariant subspaces. *IEEE Trans. Aut. Control*, **AC-27**, 1071–1085.
- Wong, K.-T. (1974). The eigenvalue problem $T\dot{x} + Sx = J$. *Diff. Equat.* **16**, 270–280.
- Wonham, W. M. (1979). *Linear Multivariable Control: a Geometric Approach*. Springer, New York.

Brief Paper

Kalman Smoothing via Auxiliary Outputs*

JOHN KORMYLO†

Key Words—Kalman filtering, smoothing, splines, state estimation, stochastic systems.

Abstract—A class of computationally efficient, fixed interval, discrete-time Kalman smoothers is discussed, of which Bierman's smoother is a special case. These smoothers obtain estimates of the complete state vector without having to store the error covariance matrices. In particular we consider the smoothing problem when the transition matrix is singular but the system is still completely reachable. Smoothing splines are used as an example.

1. Introduction

NUMEROUS fixed-interval discrete-time smoothing algorithms have appeared in the literature. Meditch (1973) and Kailath (1974, 1975) have surveyed this area, and Bierman (1973, 1977, 1982) made numerous contributions to it. A major problem with most algorithms is that the complete set of error covariance matrices must be stored between the forward and reverse passes. Bierman's most recent smoother (Bierman, 1982) avoids this problem in a stable way, but is constrained to models where the transition matrix is invertible.

This paper shows how an entire class of smoothers can be developed for completely reachable systems using the concept of auxiliary outputs. These smoothers do not require storage of the error covariance matrices; in fact, Bierman's smoother is a special case of this type of smoother. A second type of auxiliary output can be used when the transition matrix is singular, yet retaining the advantages of the Bierman smoother.

2. Problem formulation and notation

Consider a standard discrete-time linear dynamic model of the form

$$\mathbf{x}(k+1) = \Phi \mathbf{x}(k) + \Gamma \mathbf{w}(k) \quad \forall k = 0, 1, \dots, N-1 \quad (1)$$

with discrete observations

$$\mathbf{z}(k) = H \mathbf{x}(k) + \mathbf{m}(k) \quad \forall k = 1, \dots, N. \quad (2)$$

The usual statistical assumptions for Kalman filtering are made:

$$E\{\mathbf{w}(i)\mathbf{w}'(j)\} = Q\delta(i-j) \quad (3)$$

$$E\{\mathbf{m}(i)\mathbf{m}'(j)\} = R\delta(i-j) \quad (4)$$

$$E\{\mathbf{w}(k)\mathbf{x}'(0)\} = \mathbf{0} \quad \forall k \geq 0 \quad (5)$$

$$E\{\mathbf{m}(k)\mathbf{x}'(0)\} = \mathbf{0} \quad \forall k \geq 0 \quad (6)$$

and, for the sake of simplicity,

$$E\{\mathbf{w}(i)\mathbf{m}'(j)\} = \mathbf{0} \quad \forall i, j. \quad (7)$$

While matrices Φ , Γ , Q , and R are not explicitly denoted as

time varying, the following results are valid for the time varying case as well.

The predictor-corrector form of the Kalman filter under these assumptions is given by

$$\hat{\mathbf{x}}(k+1|k) = \Phi \hat{\mathbf{x}}(k|k) \quad (8)$$

$$P(k+1|k) = \Phi P(k|k)\Phi' + \Gamma Q \Gamma' \quad (9)$$

for the predictor stage, and

$$\hat{\mathbf{x}}(k|k-1) = \hat{\mathbf{x}}(k) - H \hat{\mathbf{x}}(k|k-1) \quad (10)$$

$$V(k) = H P(k|k-1) H' + R \quad (11)$$

$$K(k) = P(k|k-1) H' V^{-1}(k) \quad (12)$$

$$\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1) + K(k) \hat{\mathbf{x}}(k|k-1) \quad (13)$$

and

$$P(k|k) = [I - K(k)H] P(k|k-1) \quad (14)$$

for the corrector stage.

The modified Bryson-Fraser smoother (Bierman, 1973) computes fixed-interval estimates $\hat{\mathbf{x}}(k|N) \approx E(\mathbf{x}(k)|\mathbf{z}(1), \dots, \mathbf{z}(N))$ using

$$\hat{\mathbf{x}}(k|N) = \hat{\mathbf{x}}(k|k-1) + P(k|k-1) \mathbf{r}(k|N) \quad (15)$$

where $\mathbf{r}(k|N)$ is the adjoint state vector and is computed recursively in the reverse direction using

$$\mathbf{r}(k|N) = [I - K(k)H]' \Phi' \mathbf{r}(k+1|N) + H' V^{-1}(k) \hat{\mathbf{x}}(k|k-1) \quad (16)$$

starting from $\mathbf{r}(N+1|N) = \mathbf{0}$.

The procedure for obtaining smoothed state estimates using the modified Bryson-Fraser smoother is to perform Kalman filtering (storing the results) then recursively compute (16) in the reverse direction.

Quantities stored: $\hat{\mathbf{x}}(k+1|k)$, $P(k+1|k)$, $K(k)$ and $V^{-1}(k) \hat{\mathbf{x}}(k|k-1)$. Total = $(n + n^2 + nm + m)N$ words where n is the dimension of \mathbf{x} , m is the dimension of \mathbf{z} , and N is the number of samples for fixed interval smoothing.

Finally, using Mendel's optimal smoother (Mendel, 1977; Mendel and Kormylo, 1978) one can compute fixed-interval estimates for driving noise $\mathbf{w}(k)$ using

$$\hat{\mathbf{w}}(k|N) = Q \Gamma' \mathbf{r}(k+1|N), \quad (17)$$

where $\mathbf{r}(k|N)$ is defined as before. Note that only $K(k)$ and $V^{-1}(k) \hat{\mathbf{x}}(k|k-1)$ need be stored for these estimates.

These estimators will be used in the following derivations.

3. Auxiliary outputs

Consider an auxiliary output, $\mathbf{y}(k)$, of the form

$$\mathbf{y}(k) = A \mathbf{x}(k) + \mathbf{w}(k) \quad (18)$$

for some matrix A , so that substituting for $\mathbf{w}(k)$ in (1) yields

$$\mathbf{x}(k+1) = [\Phi - \Gamma A] \mathbf{x}(k) + \Gamma \mathbf{y}(k). \quad (19)$$

If $\{\Phi, \Gamma\}$ describes a completely reachable system, then there exists a matrix A such that

$$G = [\Phi - \Gamma A]^{-1} \quad (20)$$

* Received 8 December 1989; revised 6 June 1990; received in final form 19 July 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor T. Basar under the direction of Editor H. Kwakernaak.

† Exxon Production Research Co., PO Box 2189, Houston, TX 77252-2189, U.S.A.

exists and is stable. Since the characteristic values of a completely controllable system can be arbitrarily located (Kwakernaak and Sivan, 1972) and since the definition of controllability is mathematically equivalent to reachability, one can choose A so as to force all the characteristic values of $\Phi - \Gamma A$ to be outside the unit circle, in which case G will exist and have all of its characteristic values inside the unit circle.

From (19) and (20) we see that

$$\mathbf{x}(k) = G\mathbf{x}(k+1) - G\Gamma\mathbf{y}(k), \quad (21)$$

and taking the conditional expectation given $\mathbf{z}(1), \dots, \mathbf{z}(N)$, we obtain our desired result

$$\hat{\mathbf{x}}(k|N) = G\hat{\mathbf{x}}(k+1|N) - G\Gamma\hat{\mathbf{y}}(k|N). \quad (22)$$

The importance of (22) is that given $\hat{\mathbf{x}}(N|N)$ and smoothed estimates for auxiliary output $\mathbf{y}(k)$, one can recursively compute smoothed estimates for the state vector $\mathbf{x}(k)$. Since the dimension of $\mathbf{y}(k)$ is generally much less than $\mathbf{x}(k)$, computing estimates $\hat{\mathbf{y}}(k|N)$ will require much less storage between the two passes, as will now be shown.

Taking the conditional expectation given $\mathbf{z}(1), \dots, \mathbf{z}(N)$ of both sides of (18), we obtain

$$\hat{\mathbf{y}}(k|N) = A\hat{\mathbf{x}}(k|N) + \hat{\mathbf{w}}(k|N). \quad (23)$$

Substituting for $\hat{\mathbf{x}}(k|N)$ using (15) and for $\hat{\mathbf{w}}(k|N)$ using (17), we see that

$$\begin{aligned} \hat{\mathbf{y}}(k|N) &= A\hat{\mathbf{x}}(k|k-1) + AP(k|k-1)\mathbf{r}(k|N) \\ &\quad + Q\Gamma'\mathbf{r}(k+1|N). \end{aligned} \quad (24)$$

Also, from (12), (14), and (16) we see that

$$\begin{aligned} P(k|k-1)\mathbf{r}(k|N) &= P(k|k)\Phi'P(k+1|N) \\ &\quad + K(k)\hat{\mathbf{z}}(k|k-1) \end{aligned} \quad (25)$$

so that (24) can be reduced to the form

$$\begin{aligned} \hat{\mathbf{y}}(k|N) &= A\hat{\mathbf{x}}(k|k-1) + [AP(k|k)\Phi' + Q\Gamma']\mathbf{r}(k+1|N) \\ &\quad + AK(k)\hat{\mathbf{z}}(k|k-1) \end{aligned} \quad (26)$$

or, using (13),

$$\hat{\mathbf{y}}(k|N) = A\hat{\mathbf{x}}(k|k) + [AP(k|k)\Phi' + Q\Gamma']\mathbf{r}(k+1|N). \quad (27)$$

To obtain fixed interval smoothed state estimates one performs Kalman filtering (storing the results) then computes $\mathbf{r}(k|N)$ and $\hat{\mathbf{x}}(k|N)$ recursively in the reverse direction using (14), (22), and (27), starting from $\mathbf{r}(N+1|N) = \mathbf{0}$ and $\hat{\mathbf{x}}(N|N)$. Since the dimension of $\mathbf{y}(k)$ is generally much less than the dimension of $\mathbf{x}(k)$, this requires far less memory than the modified Bryson-Fraser smoother.

Quantities stored: $\hat{\mathbf{y}}(k|k)$, A , $AP(k|k)$, $K(k)$ and $V^{-1}(k)\hat{\mathbf{z}}(k|k-1)$. Total = $(l + 2nl + nm + m)N$ words where n is the dimension of \mathbf{x} , m is the dimension of \mathbf{z} , l is the dimension of \mathbf{w} , and N is the number of samples for fixed interval smoothing.

Obviously, this smoother requires less storage when $l \ll (n(n+1)/2n+1) = n/2$.

4. Bierman's smoother

For the special case when

$$A = -Q[\Phi^{-1}\Gamma']^{-1}(k|k) \quad (28)$$

we satisfy

$$AP(k|k)\Phi' + Q\Gamma' = \mathbf{0}, \quad (29)$$

in which case from (27) we see that our smoother does not require the adjoint state vector. The reason for this simplification is that, as one can easily show,

$$E\{\hat{\mathbf{y}}(k|k)\hat{\mathbf{z}}(j|j-1)\} = \mathbf{0} \quad \forall j \geq k \quad (30)$$

when (29) is satisfied, and therefore

$$\hat{\mathbf{y}}(k|N) = \hat{\mathbf{y}}(k|k) = A\hat{\mathbf{x}}(k|k). \quad (31)$$

This can produce considerable savings in storage, as one no longer need store the Kalman gains or innovations. The only disadvantage is that it requires the transition matrix to be invertible.

Quantities stored: $\hat{\mathbf{y}}(k|N)$ and A . Total = $(l + nl)N$ words where n is the dimension of \mathbf{x} , l is the dimension of \mathbf{w} , and N is the number of samples for fixed interval smoothing.

In addition, one can easily show that

$$\begin{aligned} [\Phi - \Gamma A]P(k|k)\Phi'P^{-1}(k+1|k) \\ = I - \Gamma[AP(k|k)\Phi' + Q\Gamma]P^{-1}(k+1|k) \end{aligned} \quad (32)$$

so that when (29) is satisfied we have

$$G = P(k|k)\Phi'P^{-1}(k+1|k) \quad (33)$$

which is the well known smoothing gain function (Bierman, 1977).

Also, one can use the matrix inversion lemma to show that

$$G = [I + \Phi^{-1}\Gamma(I - A\Phi^{-1}\Gamma)^{-1}A]\Phi^{-1} \quad (34)$$

and therefore

$$G\Gamma = \Phi^{-1}\Gamma(I - A\Phi^{-1}\Gamma)^{-1} \quad (35)$$

We can now rewrite (22) as

$$\begin{aligned} \hat{\mathbf{x}}(k|N) &= \Phi^{-1}\hat{\mathbf{x}}(k+1|N) \\ &\quad - \Phi^{-1}\Gamma(I - A\Phi^{-1}\Gamma)^{-1}[\hat{\mathbf{y}}(k|N) \\ &\quad - A\Phi^{-1}\hat{\mathbf{x}}(k+1|N)] \end{aligned} \quad (36)$$

which is useful when $\Phi^{-1}\Gamma$ is time invariant and can therefore be pre-computed.

5. Singular transition matrix

Let us now consider an auxiliary output, $\mathbf{y}(k)$, of the form

$$\mathbf{y}(k) = A\mathbf{x}(k) + B\mathbf{x}(k+1) + \mathbf{w}(k) \quad (37)$$

for some matrices A and B . Substituting for $\mathbf{w}(k)$ in (1) we now obtain

$$[I + \Gamma B]\mathbf{x}(k+1) = [\Phi - \Gamma A]\mathbf{x}(k) + \Gamma\mathbf{y}(k) \quad (38)$$

and therefore

$$\mathbf{x}(k) = [\Phi - \Gamma A]^{-1}[I + \Gamma B]\mathbf{x}(k+1) - [\Phi - \Gamma A]^{-1}\Gamma\mathbf{y}(k). \quad (39)$$

Similarly, we now have

$$\begin{aligned} \hat{\mathbf{y}}(k|N) &= A\hat{\mathbf{x}}(k|k) + B\hat{\mathbf{x}}(k+1|k) \\ &\quad + [AP(k|k)\Phi' + BP(k+1|k) + Q\Gamma']\mathbf{r}(k+1|N) \end{aligned} \quad (40)$$

in place of (27).

When Φ is singular, there exists some non-trivial matrix C such that

$$\Phi C = \mathbf{0} \quad (41)$$

where the rank of C plus the rank of Φ equals the dimension of \mathbf{x} . If we choose A using

$$A = \alpha C'P^{-1}(k|k) \quad (42)$$

for some scalar $\alpha \neq 0$, and choose B using

$$B = -Q\Gamma'P^{-1}(k+1|k) \quad (43)$$

then from (40) we now have

$$\hat{\mathbf{y}}(k|N) = \hat{\mathbf{y}}(k|k) = A\hat{\mathbf{x}}(k|k) + B\hat{\mathbf{x}}(k+1|k) \quad (44)$$

and $[\Phi - \Gamma A]^{-1}$ exists. One can also show that

$$[\Phi - \Gamma A]^{-1}[I + \Gamma B] = P(k|k)\Phi'P^{-1}(k+1|k) \quad (45)$$

which is our old friend, the smoothing gain.

Quantities stored: $\hat{y}(k|N)$, A , and B . Total = $(l + 2nl)N$ words where n is the dimension of x , l is the dimension of w , and N is the number of samples for fixed interval smoothing

Note: the only criterion for choosing α is that the condition number of $[\Phi - \Gamma A]$ should be as low as possible. Also, when $\Gamma' \Phi = 0$ (as is often the case) the solution to (43) is given by $B = -(\Gamma' \Gamma)^{-1} \Gamma'$, in which case $B \hat{x}(k+1|k) = 0$.

6. Example. Cubic smoothing splines

Cubic smoothing splines are found by fitting data points to a series of cubic polynomials constrained to be continuous up to the second derivative. This can be formulated into a Kalman smoothing problem by defining the state vector to consist of the polynomial coefficients so that the observation matrix is given by

$$H(k) = \begin{bmatrix} 1 & (t_i - t_k) & (t_i - t_k)^2 & (t_i - t_k)^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (t_j - t_k) & (t_j - t_k)^2 & (t_j - t_k)^3 \end{bmatrix} \quad (46)$$

where t_k are the knot locations and the t_i are the observation times, where

$$\{t_1, \dots, t_j\} \subset [t_k, t_{k+1}) \quad (47)$$

If we model the third derivative as a random sequence with a very large variance, from the continuity constraints we obtain

$$\Phi = \begin{bmatrix} 1 & (t_{k+1} - t_k) & (t_{k+1} - t_k)^2 & (t_{k+1} - t_k)^3 \\ 0 & 1 & 2(t_{k+1} - t_k) & 3(t_{k+1} - t_k)^2 \\ 0 & 0 & 1 & 3(t_{k+1} - t_k) \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$\Gamma = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad (48)$$

One can easily show that $\Phi C = 0$ when

$$C = [(t_{k+1} - t_k)^3 \quad -3(t_{k+1} - t_k)^2 \quad 3(t_{k+1} - t_k) \quad -1]'. \quad (49)$$

and the value of α which minimizes the condition number is given by

$$\alpha = -1/C' P^{-1}(k|k) C \quad (50)$$

Also, in this case notice that $\Gamma' \Phi = 0$, so that we need not store the B matrices.

On the other hand, if we model the third derivative as a random walk, we have

$$\Phi = \begin{bmatrix} 1 & (t_{k+1} - t_k) & (t_{k+1} - t_k)^2 & (t_{k+1} - t_k)^3 \\ 0 & 1 & 2(t_{k+1} - t_k) & 3(t_{k+1} - t_k)^2 \\ 0 & 0 & 1 & 3(t_{k+1} - t_k) \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (51)$$

which is obviously invertible. For this modeling assumption we can use the standard Bierman Smoother. The difference between the two models will only appear when Q is small.

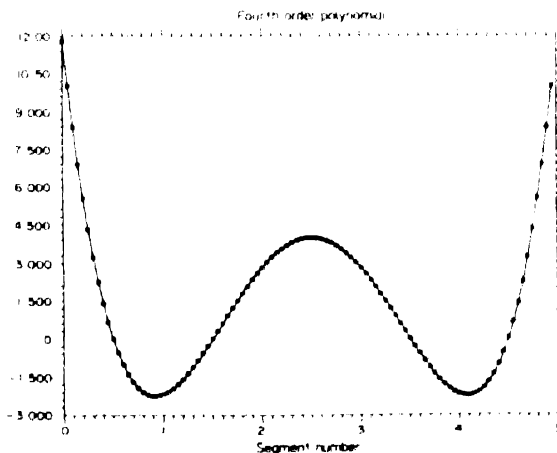


FIG. 1. Cubic spline fit to fourth order polynomial (high SNR).

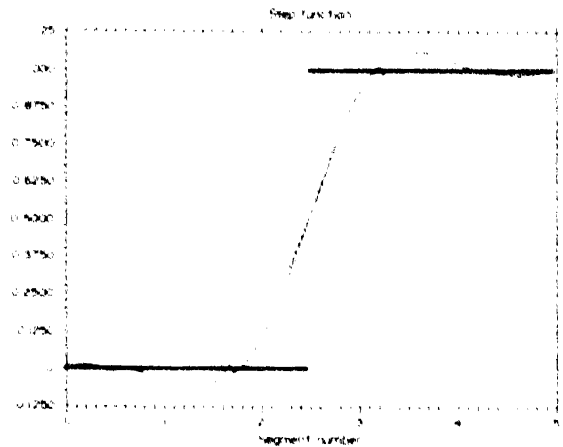


FIG. 2. Cubic spline fit to step function (low SNR)

The first model will attempt to keep the third derivative small, whereas the second model will attempt to keep the discontinuities in the third derivative small.

Figure 1 depicts the results of using both of the above models to fit the fourth order polynomial

$$z(t) = (t - 0.5)(t - 1.5)(t - 3.5)(t - 4.5) \quad (52)$$

using $Q/R = 10,000$. The circles denote the observations, the solid line shows the fit using random third derivatives, and the dashed line (hidden under the solid line) shows the fit using a random walk model. The RMS errors for the two models were 0.020295 and 0.020311, respectively. This demonstrates that both methods can produce good estimates and are virtually indistinguishable for large values of Q .

Figure 2 depicts the results of using both of the above models to fit a step function using $Q/R = 1$. The effect of the small Q is to dampen the oscillations of the resulting splines. The circles denote the observations, the solid line shows the fit using random third derivatives, and the dashed line shows the fit using a random walk model. The RMS errors for the two models were 0.139 and 0.152, respectively. This demonstrates that the two methods can produce slightly different estimates due to the differences in the modeling assumptions.

7. Conclusion

For completely reachable systems, it is possible to obtain smoothed state vector estimates by constructing auxiliary outputs and inverting the resulting system. Bierman's smoother is a special case which does not require using the adjoint state vector. When the transition matrix is singular, a second type of auxiliary output can be used with all the advantages of the Bierman smoother.

References

- Bierman, G. J. (1973). Fixed interval smoothing with discrete measurements. *Int. J. Control*, **18**, 65-75.
- Bierman, G. J. (1977). *Factorization Methods for Discrete Sequential Estimation*. Academic Press, New York.
- Bierman, G. J. (1982). A new computationally efficient fixed-interval discrete-time smoother. *Proc. Sixth IFAC Symp. on Identification and System Parameter Estimation*, Arlington, Virginia, pp. 317-321.
- Kailath, T. (1974). A view of three decades of linear filtering theory. *IEEE Trans. Inf. Theory*, **IT-20**, 146-181.
- Kailath, T. (1975). Supplement to a survey of data smoothing. *Automatica*, **11**, 109-111.
- Kwakernaak, H. and R. Sivan (1972). *Linear Optimal Control Systems*. Wiley-Interscience, New York.
- Meditch, J. S. (1973). A survey of data smoothing. *Automatica*, **9**, 151-162.
- Mendel, J. M. (1977). White noise estimators for seismic data processing in oil exploration. *IEEE Trans. Aut. Control*, **AC-22**, 694-706.
- Mendel, J. M. and J. J. Kormylo (1978). New fast optimal white-noise estimators for deconvolution. *IEEE Trans. Geosci. Electron.*, **GE-16**, 80-84.

Brief Paper

Numerical Computation of Decentralized Fixed Modes*

R. V. PATEL† and P. MISRA‡

Abstract—In this paper, we use an algebraic characterization of “fixed modes” of a decentralized linear multivariable system to show that the fixed modes are related to the “blocking zeros” of certain subsystems derived from the given decentralized system. A numerical algorithm is then presented which enables us to compute the fixed modes in a reliable manner. Examples are provided to illustrate the main results of the paper.

1. Introduction

IN RECENT years there has been considerable interest in the study of decentralized control of large scale linear multivariable systems such as those which arise in developing control strategies for large flexible space structures, e.g. West-Vukovich *et al.*, 1984 or multi-machine power systems (e.g. Davison and Tripathi, 1978). The decentralized structure of these systems is a consequence of the constraints that are imposed on the information flow within the system, usually because of the locations of various sensors and actuators. By judiciously locating these sensors and actuators, a structure can be chosen for a decentralized controller which makes it considerably simpler to implement than a “centralized” controller.

The structure of a decentralized controller is an important issue in the control of large-scale systems. This is because of the existence of “decentralized fixed modes” (d.f.m.s) (e.g. Wang and Davison, 1973; Anderson and Clements, 1981; Armentano and Singh, 1982; Corfmat and Morse, 1976; West-Vukovich *et al.*, 1984). D.f.m.s are those modes of the system which are invariant under the implementation of all decentralized controllers having a particular structure. Therefore, if a d.f.m. corresponding to a particular decentralized structure is unstable or has other undesirable characteristics, the decentralized controller will not be able to remedy the situation. One aspect of the design problem, therefore, is to develop methods of determining a structure for a decentralized controller such that there are no d.f.m.s or no undesirable d.f.m.s. Consequently, it is of interest to investigate the conditions under which these modes occur, and develop a numerically efficient and reliable method for computing them.

Several characterizations of d.f.m.s have been obtained in recent years, e.g. (e.g. Misra and Patel, 1986; Davison and Wang, 1985; Tarokh, 1985; Patel and Misra, 1984; Davison and Özgüner, 1983; Seraji, 1982; Anderson, 1982; Anderson and Clements, 1981; Corfmat and Morse, 1976; Wang and Davison, 1973). Some of these references provide charac-

terizations in terms of transmission zeros of certain “sub-systems” of the given system. The determination of d.f.m.s by these approaches can be computationally expensive for systems having high order and/or a large number of “stations”, since many transmission zero computation tests would be required. Anderson (1982) gives a transfer function characterization. However, the characterization does not provide an efficient and numerically reliable method by which d.f.m.s may be computed. Anderson and Clements (1981) give an algebraic characterization which provides valuable insight into the properties of d.f.m.s and conditions under which they occur. The characterization requires the partitioning of the set of stations into two disjoint subsets and involves a rank test, but as will be discussed later, a direct application of the result to find d.f.m.s can be computationally expensive.

One of the most straightforward ways of computing d.f.m.s is the method suggested by Wang and Davison (1973). This gives the fixed modes as those eigenvalues of the state matrix which are unaltered when random decentralized feedback is applied. However as pointed out in Misra and Patel (1986) and Vaz and Davison (1989), the method based on eigenvalue computation can be numerically unreliable. This is explained further in Section 5.

In this paper, we relate the concept of “blocking zeros” (Patel, 1986) of a linear multivariable system to the fixed modes of decentralized systems. It is shown how such a characterization leads to a numerically reliable algorithm for computing d.f.m.s.

2. Preliminaries

Definition 1. A linear time-invariant multivariable system described by

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \sum_{i=1}^N \mathbf{B}_i \mathbf{u}_i(t) \quad (2.1a)$$

$$\mathbf{y}_i(t) = \mathbf{C}_i \mathbf{x}(t), \quad i = 1, \dots, N \quad (2.1b)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$, $\mathbf{u}_i(t) \in \mathbb{R}^{m_i}$, $\mathbf{y}_i(t) \in \mathbb{R}^{p_i}$, $i = 1, \dots, N$, is called an “ N station decentralized system”.

Definition 2. Given the system (2.1), if we define a set \mathbf{K} of block-diagonal matrices \mathbf{K} as

$$\mathbf{K} = \{\mathbf{K} \mid \mathbf{K} = \text{block diag}(\mathbf{K}_1, \dots, \mathbf{K}_N), \mathbf{K}_i \in \mathbb{R}^{m_i \times p_i}\} \quad (2.2)$$

then the set of d.f.m.s of (2.1) with respect to \mathbf{K} is defined as

$$\Lambda(\mathbf{A}, \mathbf{B}_i, \mathbf{C}_i, \mathbf{K}) = \bigcap_{\mathbf{K} \in \mathbf{K}} \sigma\left(\mathbf{A} + \sum_{i=1}^N \mathbf{B}_i \mathbf{K}_i \mathbf{C}_i\right) \quad (2.3)$$

where $\sigma(\cdot)$ denotes the set of eigenvalues of the matrix (\cdot) .

Theorem 1. A scalar $\lambda \in \sigma(\mathbf{A})$ is a d.f.m. of the system described by (2.1) if and only if for *some* partition of the set $\Omega = \{1, \dots, N\}$ into disjoint subsets $\Omega_k = \{i_1, \dots, i_k\}$ and $\Omega_0 = \{i_{k+1}, \dots, i_N\}$,

$$\text{rank} \begin{bmatrix} \lambda \mathbf{I}_n - \mathbf{A} & \mathbf{B}_{i_1} & \dots & \mathbf{B}_{i_k} \\ \mathbf{C}_{i_{k+1}} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{i_N} & 0 & \dots & 0 \end{bmatrix} < n. \quad (2.4)$$

* Received 2 October 1986; revised 1 December 1988; received in final form 18 June 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor T. Başar under the direction of Editor A. P. Sage.

† Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada H3G 1M8. Author to whom all correspondence should be addressed.

‡ Department of Electrical Engineering, Wright State University, Dayton, OH 45435, U.S.A. and Concordia University, Montreal, Canada.

Proof. See Anderson and Clements (1981).

Remarks. Since, $\lambda \in \sigma(A)$, all "possible" candidates for d.f.ms are known *a priori*. Further, since the partition in (2.4) is disjoint, any mode which is controllable and observable from one or more stations *cannot* be a d.f.m. and these modes can be eliminated from consideration.

Theorem 2. Any multi-input, multi-output, single-station system (A, B, C) can be reduced by means of an orthogonal state coordinate transformation T to a condensed form $(\tilde{F}, \tilde{G}, \tilde{H})$ called a "block upper Hessenberg form" (BUHF) such that $\tilde{F} = T^T A T$, $\tilde{G} = T^T B$ and $\tilde{H} = C T$, with

$$\tilde{F} = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1,k} & F_{1,k+1} \\ F_{21} & F_{22} & \cdots & F_{2,k} & F_{2,k+1} \\ 0 & F_{32} & \cdots & F_{3,k} & F_{3,k+1} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & F_{k,k} & F_{k,k+1} \\ 0 & 0 & \cdots & 0 & F_{k+1,k+1} \end{bmatrix},$$
$$\tilde{G} = \begin{bmatrix} G_1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \quad \tilde{H} = [H_1 \ H_2 \ \cdots \ H_k \ H_{k+1}] \quad (2.5)$$

where $F_{ij} \in \mathbb{R}^{l_i \times (l_j+1)}$, $H_i \in \mathbb{R}^{p \times l_i}$ and $G_i \in \mathbb{R}^{m \times 1}$. The integers l_i , $i = 0, \dots, k$ are defined as follows: $l_0 = \text{rank}(B)$, $l_i = \text{rank}(F_{i+1,:})$, $i = 1, \dots, k$ and $\sum_{i=0}^k l_i = \mu$ where μ is the dimension of the controllable subsystem; $\mu = n$ if the system is controllable and $\mu < n$ if the system is uncontrollable.

Proof. The proof is by construction and can be found in Patel (1981), Paige (1981) and Van Dooren (1981). Reduction of a system to a block Hessenberg form has also been reported in several other publications (e.g. Nour-Eldin, 1977; Tse *et al.*, 1978; Konstantinov *et al.*, 1981), although non-orthogonal transformations have been used in some cases.

Remarks. A similar result can be stated for reducing the triple to a "block lower Hessenberg form" (BLHF). Using Theorem 2, we can easily obtain a minimal order subsystem from the given triple (A, B, C) by first determining the controllable subsystem $(A^{(c)}, B^{(c)}, C^{(c)})$ and then computing the observable subsystem $(A^{(o)}, B^{(o)}, C^{(o)})$ of $(A^{(c)}, B^{(c)}, C^{(c)})$.

Fact 1. The system (A, B, C) can be reduced by means of a unitary state coordinate transformation U to a condensed form $(\tilde{F}, \tilde{G}, \tilde{H}) = (U^H A U, U^H B, C U)$ where U^H denotes the conjugate transpose of U , such that \tilde{F} is in an upper Schur form (USF). In this condensed form, all eigenvalues of A appear along the diagonal of \tilde{F} as real or complex scalars, and can be made to appear along the diagonal of \tilde{F} in any desired order by appropriate choice of U (Golub and Van Loan, 1989).

Fact 2. The system (A, B, C) can be reduced by means of an orthogonal state coordinate transformation T to a condensed form $(\tilde{F}, \tilde{G}, \tilde{H}) = (T^T A T, T^T B, C T)$ such that \tilde{F} is in real Schur form (RSF). The eigenvalues of A appear along the diagonal of \tilde{F} with real eigenvalues as scalars and complex-conjugate pairs of eigenvalues as 2×2 blocks. These scalars and 2×2 blocks can be arranged in any desired order by appropriate choice of T (Golub and Van Loan, 1989).

Definition 3. A scalar $\lambda \in \mathbb{C}$ is a "blocking zero" (Patel, 1986) of the system (A, B, C) with A a cyclic matrix, if $C \text{adj}(\lambda I_n - A)B = 0$ where $\text{adj}(\cdot)$ denotes the adjoint of

the matrix (\cdot) . If $\lambda \notin \sigma(A)$, then λ is a blocking zero of (A, B, C) if $C(\lambda I_n - A)^{-1}B = 0$.

3. Characterization and computation of d.f.ms

In this section, we will investigate the conditions under which the inequality in (2.4) is satisfied. It will be shown later that these conditions can be easily used to derive an efficient and numerically reliable algorithm to compute the d.f.ms of (2.1).

Consider a system (F, G, H) with $F \in \mathbb{R}^{n \times n}$, $G \in \mathbb{R}^{n \times m}$ and $H \in \mathbb{R}^{p \times n}$. Assume that F is a cyclic matrix with an eigenvalue λ of multiplicity r . Also assume without loss of generality that the matrix F is in USF (see Fact 1). The system (F, G, H) therefore has the following structure:

$$F = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1,r} & F_{1,r+1} \\ 0^T & \lambda & \cdots & f_{2,r} & f_{2,r+1} \\ 0^T & 0 & \cdots & f_{3,r} & f_{3,r+1} \\ \vdots & \vdots & & \vdots & \vdots \\ 0^T & 0 & \cdots & \lambda & f_{r,r+1} \\ 0^T & 0 & \cdots & 0 & \lambda \end{bmatrix},$$
$$G = \begin{bmatrix} G_1 \\ G_2^T \\ \vdots \\ G_r^T \\ G_{r+1}^T \end{bmatrix}, \quad H = [H_1 \ h_2 \ \cdots \ h_r \ h_{r+1}] \quad (3.1)$$

Since $\lambda \notin \sigma(F_{11})$ and F is a cyclic matrix, $\text{rank}(\lambda I - F) = n - 1$, which implies that $f_{i,r+1} \neq 0$, $i = 2, \dots, r$. We can now state the following result:

Theorem 3. Let the system (F, G, H) defined above have an uncontrollable and unobservable mode at λ . Also, assume (without loss of generality) that the eigenvalue λ corresponding to this mode is in the (n, n) th position of F . Then $\text{rank} \begin{bmatrix} \lambda I_n - F & G \\ H & 0 \end{bmatrix} < n$ if and only if λ is a blocking zero of the system $(\tilde{F}, \tilde{G}, \tilde{H})$ where

$$\tilde{F} = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1,r+1} \\ 0^T & \lambda + f_{2,r} & \cdots & f_{2,r+1} \\ 0^T & \vdots & & \vdots \\ 0^T & 0 & \cdots & \lambda + f_{r,r+1} \end{bmatrix},$$
$$\tilde{G} = \begin{bmatrix} G_1 \\ G_2^T \\ \vdots \\ G_r^T \end{bmatrix}, \quad \tilde{H} = [H_1 \ h_2 \ \cdots \ h_{r+1}] \quad (3.2)$$

Proof. See Appendix.

The above result can be easily applied in characterizing d.f.ms of (2.1). Once the set Ω has been partitioned into disjoint subsets Ω_i and Ω_0 , the matrix G will be the partitioned matrix $[G_0 \ G_1 \ \cdots \ G_n]$ and the matrix H will be the partitioned matrix $[H_1^T \ H_2^T \ \cdots \ H_n^T]^T$.

One way to obtain the partitions Ω_i and Ω_0 is to find all the stations from which λ is uncontrollable and all the stations from which it is unobservable. This can be done merely by inspection once A has been reduced to its USF F with λ at position (n, n) and $(1, 1)$ to check for uncontrollability and unobservability respectively. For complex-conjugate pairs of eigenvalues, we can avoid the use of complex arithmetic by reducing A to an RSF, with the corresponding 2×2 blocks at the bottom right and top left corners to check for uncontrollability and unobservability respectively. However, the information obtained from the above inspection may not necessarily give a disjoint partition i.e. $\Omega_i \cap \Omega_0 \neq \emptyset$ where \emptyset is the null set.

If there is a station γ such that $\gamma \in \Omega_i \cap \Omega_0$, then for (2.4) to hold, λ must necessarily be uncontrollable and

unobservable from the y th station. Let $\Psi = \Omega_i \cap \Omega_0 \neq \emptyset$. To verify the rank condition in (2.4), it is necessary to assign each element of Ψ to Ω_i or Ω_0 such that $\Omega_i \cap \Omega_0 = \emptyset$ and at the same time, the partition should be such that if λ is a d.f.m. of (2.1), then the rank condition in (2.4) is satisfied.

The problem of computing the d.f.ms of the system described by (2.1) can be divided into two smaller problems. (1) Obtaining a set $\hat{\Lambda}$ such that the set of fixed modes $\Lambda \subseteq \hat{\Lambda} \subseteq \sigma(A)$. The set $\hat{\Lambda}$ consists of all possible candidates for d.f.ms; and (2) obtaining a disjoint partition of Ω (if it exists) such that the rank condition (2.4) is satisfied.

To find the set $\hat{\Lambda} \subseteq \sigma(A)$ that consists of the possible candidates for the d.f.ms, we remove all the eigenvalues of A that are controllable and observable from the same station. This can be accomplished by reductions to BUHF and BLHF using orthogonal transformations as described in Theorem 2.

Remark. All $\lambda_i \in \hat{\Lambda}$ need not be d.f.ms of the given system. The set $\hat{\Lambda}$ contains those eigenvalues of the system which are possible candidates for d.f.ms. Usually the set $\hat{\Lambda}$ is a very small subset of $\sigma(A)$.

Having obtained $\hat{\Lambda}$, we now need to examine each element of $\hat{\Lambda}$ to determine whether or not it is a d.f.m. of the system. We denote the subsystem obtained at the end of the above procedure (after the controllable and observable subsystems from all stations have been removed) by an \hat{n} th order system $(\hat{A}, \hat{B}_i, \hat{C}_i)$, $i = 1, \dots, N$. The elements of $\hat{\Lambda}$ are the eigenvalues of \hat{A} .

In order to obtain the disjoint partitions Ω_i and Ω_0 (if they exist) that also satisfy (2.4), it will be necessary to evaluate at some complex value λ several "transfer function relations" of the form

$$S_{ij} = \hat{C}_i(\lambda I_{\hat{n}} - \hat{A})^{-1} \hat{B}_j \quad (3.3)$$

where $i, j \in \Omega$ and $i \neq j$. The matrices \hat{A} , \hat{B}_i and \hat{C}_i above are obtained from the system (A, B_i, C_i) as shown below. Let

$$A = \begin{bmatrix} A_{11} & a_{12} & \dots & a_{1r} & a_{1,r+1} \\ 0^T & \lambda & \dots & a_{2r} & a_{2,r+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0^T & 0 & \dots & \lambda & a_{r,r+1} \\ 0^T & 0 & \dots & 0 & \lambda \end{bmatrix}, \quad B_i = \begin{bmatrix} B_{1i} \\ b_{2i}^T \\ \vdots \\ b_{ri}^T \\ b_{r+1,i}^T \end{bmatrix}, \quad C_i = [C_{i1} \ c_{i2} \ \dots \ c_{ir} \ c_{i,r+1}] \quad (3.4)$$

Then

$$\hat{A} = \begin{bmatrix} A_{11} & a_{12} & \dots & a_{1,r+1} \\ 0^T & \lambda + a_{22} & \dots & a_{2,r+1} \\ \vdots & \vdots & \ddots & \vdots \\ 0^T & 0 & \dots & \lambda + a_{r,r+1} \end{bmatrix}, \quad \hat{B}_i = \begin{bmatrix} B_{1i} \\ b_{2i}^T \\ \vdots \\ b_{ri}^T \\ b_{r+1,i}^T \end{bmatrix}, \quad \hat{C}_i = [C_{i1} \ c_{i2} \ \dots \ c_{ir} \ c_{i,r+1}] \quad (3.5)$$

Next, we will discuss a systematic procedure which enables us to find the required disjoint partition of Ω into Ω_i and Ω_0 .

Algorithm 1 (To find disjoint sets Ω_i and Ω_0).

- Step 1:** (a) Set $q = 1$ and $\Lambda = \emptyset$.
 (b) Transform $(\hat{A}, \hat{B}_i, \hat{C}_i)$, $i = 1, \dots, N$, such that \hat{A} is in USF with λ_q in its $(1, 1)$ position.
 (c) Form a set $\hat{\Omega}_0$ where $\hat{\Omega}_0$ contains all the stations for which the first columns of the output matrices \hat{C}_i , $i = 1, \dots, N$, are zero vectors.
- Step 2:** (a) Transform $(\hat{A}, \hat{B}_i, \hat{C}_i)$, $i = 1, \dots, N$, such that \hat{A} is in USF with λ_q in its (\hat{n}, \hat{n}) position.
 (b) Form a set $\hat{\Omega}_i$ where $\hat{\Omega}_i$ contains all the stations

for which the last rows of the input matrices \hat{B}_i , $i = 1, \dots, N$, are zero vectors.

Step 3: Let the sets $\hat{\Omega}_i$ and $\hat{\Omega}_0$ be given by $\hat{\Omega}_i = \{i_1^*, \dots, i_{k_i}^*, i_{k_i+1}, \dots, i_n\}$, $\hat{\Omega}_0 = \{i_1, \dots, i_n, i_{k_1+1}^*, \dots, i_{k_n}^*\}$, where the "asterisked" elements are the ones which correspond to the stations from where λ_q is either controllable but unobservable or uncontrollable but observable. Form $\Psi = \hat{\Omega}_i \cap \hat{\Omega}_0 = \{i_{k_1+1}^*, \dots, i_{k_n}^*\}$ and set $\hat{\Omega}_i = \{i_1^*, \dots, i_{k_i}^*\}$, $\hat{\Omega}_0 = \{i_1^*, \dots, i_{k_n}^*\}$ and $k = 1$.

Step 4: (a) Corresponding to the k -th (asterisked) station of $\hat{\Omega}_0$, form $S_{i_{k+1}, i_k} = \hat{C}_{i_{k+1}}(\lambda_q I_{\hat{n}} - \hat{A})^{-1} \hat{B}_{i_k}$, $i_k \in \hat{\Omega}_i$.

- (b) For $i \in \hat{\Omega}_i$
 (i) If $S_{i_{k+1}, i_k} \neq 0$ and $i \in \hat{\Omega}_i$, go to Step 5
 (ii) If $S_{i_{k+1}, i_k} \neq 0$ and $i \in \hat{\Omega}_0$,
 Set $\hat{\Omega}_i = \hat{\Omega}_i \cup \{i\}$ (remove i from $\hat{\Omega}_0$)
 Set $\hat{\Omega}_0 = \hat{\Omega}_0 \cup \{i\}$ (add i to $\hat{\Omega}_0$)
 Set $\Psi = \Psi \cup \{i\}$, go to Step 4(c)
 (iii) If $S_{i_{k+1}, i_k} = 0$ for all $i \in \hat{\Omega}_i$, go to Step 4(c)
 (c) If all (asterisked) stations of $\hat{\Omega}_i$ are exhausted, go to Step 6.

Else, set $k = k + 1$ and go to Step 4(a).

Step 5: There is no disjoint partition for which (2.4) is satisfied. Therefore, λ_q is not a d.f.m. Go to Step 7.

Step 6: Set $\hat{\Omega}_i = \hat{\Omega}_i$; the disjoint partition satisfying (2.4) is given by $\hat{\Omega}_0$ and $\hat{\Omega}_i$. Therefore λ_q is a d.f.m. Set $\Lambda = \Lambda \cup \lambda_q$ and go to Step 7.

Step 7: If $q = n$, stop; else set $q = q + 1$ and go to Step 1(b). At the end of Algorithm 1, the set $\Lambda(\hat{C}_i, \hat{A})$ contains all the eigenvalues of \hat{A} which are d.f.ms of (2.1).

Remarks. When A has some complex conjugate pairs of eigenvalues, we can use the reduction to RSF in Steps 1 and 2 to avoid complex arithmetic. In Step 4b, the condition in (i) implies that for at least one station that must appear in any disjoint partition of Ω , the transfer function relation evaluated at λ_q is non-zero. This, in turn, means that there is no disjoint partition of Ω that satisfies (2.4) for λ_q . The condition in (ii) corresponds to the case where the station is in both $\hat{\Omega}_i$ and $\hat{\Omega}_0$. If the condition is satisfied, then the station cannot be included in $\hat{\Omega}_i$. So we include it in $\hat{\Omega}_0$ and remove it from $\hat{\Omega}_i$ and Ψ . In order to conform with the value of k in subsequent steps, we shall assume that the operation of set addition in Step 4b (ii) corresponds to including the station i at the end of the set $\hat{\Omega}_0$. Condition (iii) does not provide any additional information about forming disjoint partitions of Ω except to say that λ_q could be a d.f.m. If Ψ is the empty set but λ_q is not a d.f.m., then condition (i) would be satisfied.

4 Discussion of the results

In this section, we will discuss various computational and numerical properties of the proposed algorithms.

4.1 General remarks. 1 Note that the computational procedure developed in this paper assumes that the state matrix is cyclic. This restricts the class of systems for which we can compute d.f.ms. However, it should be pointed out that the eigenvalue problem for non-cyclic matrices is often likely to be poorly conditioned (Stewart, 1973). Therefore, any procedure that requires knowledge of the eigenvalues of such non-cyclic state matrices will give inaccurate results.

2 Decentralization can be considered as one class of constraints that can be imposed on a feedback structure. A more general class can be defined via arbitrary constraints on the elements of feedback matrices. For such systems, the proposed method will not be suitable for computing "fixed modes" whereas the eigenvalue approach of Wang and Davison (1973) can be used.

3 The characterization of d.f.ms in terms of blocking zeros of certain subsystems derived from the given decentralized system is consistent with the characterization given by Davison and Wang (1985) of d.f.ms in terms of transmission zeros. Such characterizations provide "structural" information about the decentralized system. It should be mentioned that the characterization does not impose any restriction on the structure of the decentralized system.

Simpler results can of course be obtained when the system has certain additional properties, e.g. block diagonal structure for A , or interconnected systems with only input and output matrices in block diagonal form.

4. The algorithm for obtaining Λ , uses only orthogonal state coordinate transformations and is numerically "backward stable" (Van Dooren, 1981). This is a desirable property of the algorithm from the point of view of its application to very high order systems. For greater reliability, singular value decomposition (Patel, 1981) can be used for the reduction of the system to its BUHF.

5. The breakdown of the operations count required for the proposed algorithms is as follows.

(a) Obtaining the matrix A : This step requires several reductions to block lower and upper Hessenberg forms and involves approximately

$$\sum_{k=1}^r n_k^2 \left(3n_k + \sum_{i=k}^N (m_i + p_i) \right)$$

operations, where $n_1 = n$, the dimension of the original system; n_k , $k \geq 2$, is the dimension of the subsystem that is uncontrollable and/or unobservable from the previous $k-1$ stations; r is an integer which is the smaller of N and the minimum number of stations from which the entire system is controllable and observable.

(b) Finding the elements of the set $\hat{\Lambda}$: These elements correspond to the eigenvalues of \hat{A} and typically $\hat{n} \ll n$. Note however that since we need to evaluate the transfer function relations S_{ij} , it would be advantageous to have A in upper Hessenberg form. Reducing the entire system to upper Hessenberg form would require approximately

$$n^2 \left(\hat{n} + \sum_{i=1}^N (m_i + p_i) \right)$$

operations.

(c) Obtaining the partitions $\hat{\Omega}_i$ and $\hat{\Omega}_i$: For this step we need to rearrange the eigenvalues of \hat{A} to determine the stations from which a particular element of $\hat{\Lambda}$ is uncontrollable and/or unobservable. In this case, since the value of the shift in the QR algorithm is known (Golub and Van Loan, 1989), this requires approximately \hat{n}^2 operations for each QR step. Including the transformations on \hat{B}_i and \hat{C}_i , we need approximately

$$\hat{n}^2 \left(\hat{n} + \sum_{i=1}^N (m_i + p_i) \right)$$

operations for this step. If $\hat{\Lambda}$ contains a complex-conjugate pair of eigenvalues, then an implicit double shift can be used to get an RSF.

(d) The number of transfer function relations S_{ij} that need to be evaluated to find the disjoint partition depends on the given system data and therefore cannot be specified *a priori*. However, a reasonable operations count for each $\lambda_q \in \hat{\Lambda}$ is approximately

$$(n-1)^2 \left(n-1 + \sum_{i=1}^s m_i + \sum_{i=1}^{N-s} p_i \right)$$

where s and $N-s$ are respectively the number of elements in $\hat{\Omega}_0$ and $\hat{\Omega}_1$.

Note that in Step (a), for most practical systems, n_k , $k \geq 2$, would generally be much smaller than n , i.e. most of the modes of the system would typically be controllable and observable from many of the stations. Therefore, taking into account all the operations mentioned above and dropping the less significant terms from each step, we require approximately

$$(2 + \gamma)n^2 \left(\hat{n} + 4 + \sum_{i=1}^N (m_i + p_i) \right)$$

operations for computing all the decentralized fixed modes, where γ is the number of transfer function relations S_{ij} that we need to compute in order to find the disjoint partition of $\hat{\Omega}$. If the approach in Anderson and Clements (1981) is used,

the count will be considerably higher because it requires several rank tests on systems of order greater than n .

4.2. Illustrative example. We shall now illustrate Algorithm 1 by an example of a decentralized system with $\hat{\Omega}_r = \{4, 1, 2, 5, 6, 7\}$, $\hat{\Omega}_0 = \{1, 2, 5, 6, 7, 3\}$, $\hat{\Omega}_1 = \{4\}$ and $\hat{\Omega}_2 = \{3\}$. Therefore, using the notation in the algorithm, $s = 1$ and $v = 6$. Given below are the results at various steps of the algorithm for checking if a scalar $\lambda_q \in \hat{\Lambda}$ is a d.f.m. of the given system:

$k = 1$

Step 4a: $i_{v+1,k} = 3$. Evaluate $[S_{31}, S_{32}, S_{34}, S_{35}, S_{36}, S_{37}]$ at $\lambda_q \in \hat{\Lambda}$. Let $S_{34} = 0$ and $S_{31}, S_{35} \neq 0$.

Step 4b(i): This condition is not satisfied ($S_{34} = 0$).

Step 4b(ii): Since $S_{31}, S_{35} \neq 0$, remove stations 1 and 6 from $\hat{\Omega}_1$ and Ψ and include them in $\hat{\Omega}_0$. This gives $\hat{\Omega}_1 = \{2, 4, 5, 7\}$, $\Psi = \{2, 5, 7\}$, $\hat{\Omega}_2 = \{4\}$ and $\hat{\Omega}_0 = \{3, 1, 6\}$

$k = 2$

Step 4a: $i_{v+1,k} = 1$. Let $S_{12} \neq 0$.

Step 4b(i): This condition is not satisfied ($S_{14} = 0$).

Step 4b(ii): Since $S_{12} \neq 0$, remove station 2 from $\hat{\Omega}_1$ and include it in $\hat{\Omega}_0$. This gives $\hat{\Omega}_1 = \{4, 5, 7\}$, $\Psi = \{5, 7\}$, $\hat{\Omega}_2 = \{4\}$ and $\hat{\Omega}_0 = \{3, 1, 6, 2\}$

$k = 3$

Step 4a: $i_{v+1,k} = 6$. Let $S_{6i} = 0$ for all $i \in \hat{\Omega}_1$.

$k = 4$

Step 4a: $i_{v+1,k} = 2$. Let $S_{2i} = 0$ for all $i \in \hat{\Omega}_1$.

Step 6: λ_q is a d.f.m. and the disjoint partition $\hat{\Omega}_0 = \{1, 2, 3, 6\}$ and $\hat{\Omega}_1 = \{4, 5, 7\}$ satisfies the rank condition (2.4).

In the above example, if either S_{24} or S_{64} were non-zero, then λ_q would not be a d.f.m. If S_{65} and S_{67} were non-zero, but S_{54} and S_{74} were zero, then λ_q would be a d.f.m. with $\hat{\Omega}_0 = \{1, 2, 3, 5, 6, 7\}$ and $\hat{\Omega}_1 = \{4\}$. The algorithm ends when a disjoint partition is found or else when a conclusion is reached that no disjoint partition (for the value of λ_q under consideration) satisfying (2.4) exists.

5. Numerical examples

In this section, we consider two numerical examples to illustrate the proposed algorithms.

Example 1. Consider a 3-station decentralized system. The matrices describing the system are given in Table 1. The state matrix has eigenvalues at $\{-2, -1.5, -1.0, 3.0, 2.5, 2.0, 1.5, 1.0\}$. It is found that $\hat{\Lambda} = \{2.0\}$ i.e. only $\lambda = 2.0$ is a possible d.f.m. Next, it is found that $\lambda = 2.0$ is unobservable from stations 1 and 2 and uncontrollable from stations 1 and 3. Therefore $\hat{\Omega}_0 = \{i_1, i_2^*\}$ and $\hat{\Omega}_1 = \{i_1, i_3^*\}$. Following the steps of Algorithm 1, it is found that $S_{21} = C_{21}(\lambda_q I_{n-1} - A_{11})^{-1} B_{11}$ and $S_{31} = C_{31}(\lambda_q I_{n-1} - A_{11})^{-1} B_{11}$ are both zero matrices for $\lambda_q = 2.0$. The elements of the matrices S_{21} and S_{31} are given in Table 2. They are of the order of 10^{-16} and can be safely assumed to be zero. Therefore, we have the partition $\hat{\Omega}_1 = \{i_1, i_3\}$ and $\hat{\Omega}_0 = \{i_2\}$ which are disjoint and satisfy the condition in (2.4) i.e. $\lambda = 2.0$ is a d.f.m. of the given system.

Example 2. In this example, we will illustrate a difficulty that may be encountered in deciding whether or not a particular eigenvalue is a d.f.m. using the test proposed in Wang and Davison (1973). For instance, depending on how a "random" decentralized feedback matrix affects the closed-loop eigenvalue problem, it may not always be possible to say conclusively by inspection of the open-loop and closed-loop eigenvalues whether or not there are any d.f.ms. The difficulty is increased further when the eigenvalue problem for the open-loop state matrix is ill-conditioned. The characterization and computational algorithms presented here enable us to conclude with greater certainty if a given mode is a d.f.m. or not. The data for Example 2 is given in Table 3. For several randomly generated values of feedback matrices with their elements of the order 10^2 , it was found that certain eigenvalues of the system do not "move"

TABLE 1. VARIOUS MATRICES IN EXAMPLE 1

$A =$	$\begin{bmatrix} -2.902280128 & 4.902280128 & 3.698135576 & -1.091710896 & -1.339597088 & -3.833588112 & .1354525359 \\ 14.57598539 & 8.663631092 & 4.471485536 & -6.163631092 & -1.031929416 & 1.692145555 & -.5280631913 \\ -13.01717766 & 5.242845433 & 6.808655713 & .3658471056 & -1.250335699 & -9.401666250 & -.1838540200 \\ -9.358517860 & 9.358517860 & 9.233430521 & -1.249825322 & -1.102425449 & -10.21076863 & .9773381096 \\ 10.90157850 & 10.33803799 & 6.577811264 & -6.527468755 & -1.463336353 & -1.277505940 & 1.391388959 \\ 7.438961098 & 6.026323164 & 4.365195826 & -4.228199859 & -1.659477876 & -1.136995967 & .6866298675 \\ 4.500977560 & 7.475549489 & 4.982190956 & -3.664980257 & -2.282092104 & -2.544374131 & .4770129007 \\ 16.85292654 & 5.785447163 & 2.204831037 & -5.785447163 & -1.798583916 & 3.580616126 & 1.906247120 \end{bmatrix}$
$B_1 =$	$\begin{bmatrix} -6.288091946 & -8.384122594 & -10.48015324 & 24.06096115 & 8.869414499 & 7.808913485 & -3.144045973 \\ 6.989292416 & 9.319056555 & 11.64882069 & 16.40558839 & 19.28474045 & 8.940485304 & 2.912205174 \\ -13.54775782 & -18.06367710 & -22.57959637 & 25.20750633 & 18.08277192 & 31.47867615 & 3.494646208 \\ -5.198332933 & -6.931110578 & -8.663888222 & 23.84960895 & 19.94822168 & 17.03165000 & -6.773878912 \\ -7.648224421 & -10.19763256 & -12.74704070 & 10.15737156 & 9.939200440 & 4.700713320 & -2.599166467 \\ -5.899533404 & -7.866044538 & -9.832555673 & 19.48795695 & 26.58611759 & 31.63366095 & -2.165972056 \\ -6.288091946 & -8.384122594 & -10.48015324 & 29.74830670 & 25.04543540 & 17.08815424 & -3.186760175 \\ 1.360132475 & 1.813509966 & 2.266887458 & 17.73934984 & 17.02037591 & 14.47725433 & -3.824112210 \\ & & & & & & -2.949766702 \\ & & & & & & -3.144045973 \\ & & & & & & -13.10019155 \\ & & & & & & 2.833609322 \end{bmatrix}$
$B_2 =$	$\begin{bmatrix} -2.632199367 & -8816532905 & 2216458770 & 8816532905 & .3060021972 & -1.103299168 & 1.2230998002 \\ .6965370444 & -6965370444 & -6386855981 & .0878917469 & .0698972579 & .6507314098 & -.0120458116 \\ 1.251321667 & 7413375337 & 3.578384236 & -2660527380 & -.0994345890 & -3.687097343 & -3.700738679 \\ .9728813679 & -6824007787 & 8321248154 & .0737554812 & .0515436172 & -8059427381 & -1.726710642 \\ 6338502954 & 1.253223829 & 1.874876846 & -7779391327 & -2226901107 & -1.471703558 & -7446180000 \\ 0725179078 & -0725179078 & -4003742177 & .6048139599 & -1581549111 & .5700763364 & -1.1697021187 \\ .0952676686 & -0687101141 & .0066780601 & .0905768308 & -1072959550 & .1535121891 & 1560969942 \\ -2252498152 & .0992129486 & .1715238903 & 1417124980 & -0477814471 & -.0948070092 & .1258808196 \\ -2314675213 & 1207746598 & 2092921292 & .0112031403 & -.0280553044 & -.0858390289 & .3357822914 \\ & & & & & & -.0594501770 \end{bmatrix}$
$C_1 =$	$\begin{bmatrix} -2.632199367 & -8816532905 & 2216458770 & 8816532905 & .3060021972 & -1.103299168 & 1.2230998002 \\ .6965370444 & -6965370444 & -6386855981 & .0878917469 & .0698972579 & .6507314098 & -.0120458116 \\ 1.251321667 & 7413375337 & 3.578384236 & -2660527380 & -.0994345890 & -3.687097343 & -3.700738679 \\ .9728813679 & -6824007787 & 8321248154 & .0737554812 & .0515436172 & -8059427381 & -1.726710642 \\ 6338502954 & 1.253223829 & 1.874876846 & -7779391327 & -2226901107 & -1.471703558 & -7446180000 \\ 0725179078 & -0725179078 & -4003742177 & .6048139599 & -1581549111 & .5700763364 & -1.1697021187 \\ .0952676686 & -0687101141 & .0066780601 & .0905768308 & -1072959550 & .1535121891 & 1560969942 \\ -2252498152 & .0992129486 & .1715238903 & 1417124980 & -0477814471 & -.0948070092 & .1258808196 \\ -2314675213 & 1207746598 & 2092921292 & .0112031403 & -.0280553044 & -.0858390289 & .3357822914 \\ & & & & & & -.0594501770 \end{bmatrix}$
$C_2 =$	$\begin{bmatrix} .9728813679 & -6824007787 & 8321248154 & .0737554812 & .0515436172 & -8059427381 & -1.726710642 \\ 6338502954 & 1.253223829 & 1.874876846 & -7779391327 & -2226901107 & -1.471703558 & -7446180000 \\ 0725179078 & -0725179078 & -4003742177 & .6048139599 & -1581549111 & .5700763364 & -1.1697021187 \\ .0952676686 & -0687101141 & .0066780601 & .0905768308 & -1072959550 & .1535121891 & 1560969942 \\ -2252498152 & .0992129486 & .1715238903 & 1417124980 & -0477814471 & -.0948070092 & .1258808196 \\ -2314675213 & 1207746598 & 2092921292 & .0112031403 & -.0280553044 & -.0858390289 & .3357822914 \\ & & & & & & -.0594501770 \end{bmatrix}$
$C_3 =$	$\begin{bmatrix} .0952676686 & -0687101141 & .0066780601 & .0905768308 & -1072959550 & .1535121891 & 1560969942 \\ -2252498152 & .0992129486 & .1715238903 & 1417124980 & -0477814471 & -.0948070092 & .1258808196 \\ -2314675213 & 1207746598 & 2092921292 & .0112031403 & -.0280553044 & -.0858390289 & .3357822914 \\ & & & & & & -.0594501770 \end{bmatrix}$

† Note that the above matrices have been rounded off to 10 decimal places. A copy of the exact matrices can be obtained from the authors

TABLE 2. TRANSFER RELATIONS S_{21} AND S_{23}

$S_{21} =$	$\begin{bmatrix} -1.110223024625157d - 16 & -1.110223024625157d - 16 & -3.330669073875470d - 16 \\ -2.220446049250313d - 16 & -4.440892098500626d - 16 & -4.440892098500626d - 16 \\ 2.220446049250313d - 16 & 0.000000000000000d + 00 & 4.440892098500626d - 16 \end{bmatrix}$
$S_{23} =$	$\begin{bmatrix} -8.326672684688674d - 17 & -5.551115123125783d - 17 & -2.220446049250313d - 16 \\ -1.110223024625157d - 16 & -1.110223024625157d - 16 & 0.000000000000000d + 00 \\ 1.110223024625157d - 16 & 1.110223024625157d - 16 & 4.440892098500626d - 16 \end{bmatrix}$

appreciably. However, as the magnitude of the elements of the feedback matrices were gradually increased to 10^{10} , the eigenvalues changed completely. Table 4 compares typical values of some of the open-loop eigenvalues with those of the closed-loop eigenvalues for 2 representative sets of values of feedback gains: $k_y \sim 10^2$ and $k_y \sim 10^{10}$ where the k_y 's are the elements of the randomly generated decentralized feedback matrices.

It should be noted that the eigenvalues at -3.0 and 2.5 do not change appreciably for $k_y \sim 10^2$ but are altered completely for $k_y \sim 10^{10}$. In contrast, corresponding to the eigenvalue at -4.6 , we have eigenvalues at -4.59492 and -4.46691 for $k_y \sim 10^2$ and 10^{10} respectively. Such observations with this and several other examples suggest that some difficulties could arise in computing the set of d.f.ms using the characterization given in (2.3). Applying the algorithms proposed above, it was found conclusively that the system does not have any d.f.ms. This was further confirmed by performing the rank test (using the singular value decomposition) on the system matrix in (2.4).

A question that arises from Example 2 is what numerical or "threshold" value should be used for "zero" in the algorithms proposed in this paper. Among the numerical techniques used in the algorithms, the maximum error is accumulated in the reduction of the state matrix to an RSF and hence an error bound on this reduction (Stewart, 1973; Wilkinson, 1965; Golub and Van Loan, 1989) may be used to define a value for "zero".

6. Conclusions

In this paper, we have used a characterization of d.f.ms given by Anderson and Clements (1981) to define d.f.ms in terms of blocking zeros of certain subsystems of the given decentralized system. Based on this characterization, an efficient and reliable method has been proposed for computing d.f.ms. The computational method uses only orthogonal transformations and can be easily implemented with software available in scientific programming packages such as IMSL, EISPACK, LINPACK, etc. Extensive numerical tests carried out so far suggest that the proposed approach is numerically more reliable than existing methods for computing d.f.ms.

Acknowledgment—The authors are grateful to the reviewers for their helpful comments and suggestions. This research was supported by the Natural Sciences and Engineering Research Council of Canada under Grant A1345.

References

- Anderson, B. D. O. and D. J. Clements (1981). Algebraic characterization of fixed modes in decentralized control. *Automatica*, **17**, 703–712.
- Anderson, B. D. O. (1982). Transfer function description of decentralized fixed modes. *IEEE Trans. Aut. Control*, **AC-27**, 1176–1182.
- Armentano, V. A. and M. G. Singh (1982). A procedure to eliminate decentralized fixed modes. *IEEE Trans. Aut. Control*, **AC-27**, 258–260.
- Corfmat, J. P. and A. S. Morse (1976). Decentralized control of linear multivariable systems. *Automatica*, **12**, 479–496.
- Davison, E. J. and N. Tripathi (1978). The optimal decentralized control of a large power system: load and frequency control. *IEEE Trans. Aut. Control*, **AC-23**, 312–325.
- Davison, E. J. and Ü. Özgüner (1983). Characterization of decentralized fixed modes for interconnected systems. *Automatica*, **19**, 169–182.
- Davison, E. J. and S. H. Wang (1985). A characterization of decentralized fixed modes in terms of transmission zeros. *IEEE Trans. Aut. Control*, **AC-30**, 81–82.
- Golub, G. H. and C. F. Van Loan (1989). *Matrix Computations*, 2nd edn. Johns Hopkins University Press, Baltimore.
- Konstantinov, M., P. Petkov and N. Christov (1981). Invariants and canonical forms for linear multivariable systems under the action of orthogonal transformation groups. *Kybernetika*, **17**, 413–424.
- Misra, P. and R. V. Patel (1986). Characterization and computation of decentralized fixed modes of multivariable systems. *Proc. 1986 Amer. Control Conf.*, Seattle, 427–432.
- Nour-Eldin, H. (1977). Minimalrealisierung der Matrix-Übertragungsfunktion. *Regelungstechnik*, **25**, 82–87.
- Paige, C. C. (1981). Properties of numerical algorithms related to computing controllability. *IEEE Trans. Aut. Control*, **AC-26**, 130–138.
- Patel, R. V. (1981). Computation of minimal order state-space realizations and observability indices using orthogonal transformations. *Int. J. Control*, **33**, 227–246.
- Patel, R. V. and P. Misra (1984). A numerical test for transmission zeros with applications in characterizing decentralized fixed modes. *Proc. 23rd IEEE Conf. on Decision and Control*, Las Vegas, 1746–1751.
- Patel, R. V. (1986). On blocking zeros in linear multivariable systems. *IEEE Trans. Aut. Control*, **AC-31**, 239–241.
- Seraji, H. (1982). On fixed modes in decentralized control systems. *Int. J. Control*, **35**, 775–784.
- Stewart, G. W. (1973). *Introduction to Matrix Computations*. Academic Press, New York.
- Tarokh, M. (1985). Fixed modes in multivariable systems using constrained controllers. *Automatica*, **21**, 495–497.
- Tse, E. C. Y., J. V. Medanic and W. R. Perkins (1978). Generalized Hessenberg transformations for reduced order modelling of large-scale systems. *Int. J. Control*, **27**, 493–512.
- Van Dooren, P. (1981). The generalized eigenstructure problem in linear systems theory. *IEEE Trans. Aut. Control*, **AC-26**, 111–129.
- Vaz, A. and E. J. Davison (1989). On the quantitative characterization of approximate decentralized fixed modes using transmission zeros. *Math. Control Signals Syst.*, **2**, 287–302.
- Wang, S. H. and E. J. Davison (1973). On stabilization of decentralized control systems. *IEEE Trans. Aut. Control*, **AC-18**, 473–478.
- West-Vukovich, G. S., E. J. Davison and P. C. Hughes (1984). The decentralized control of large flexible space structures. *IEEE Trans. Aut. Control*, **AC-29**, 866–879.
- Wilkinson, J. H. (1965). *The Algebraic Eigenvalue Problem*. Oxford University Press, London.

Appendix

Proof of Theorem 3. Since λ is an uncontrollable mode of (F, G, H) , it follows from the structure of F and G that, $\mathbf{g}_{i+1} = \mathbf{0}$. Also, since λ is an unobservable mode of (F, G, H) and F is a cyclic matrix, we have

$$\text{rank} \begin{bmatrix} \lambda I_n \\ H \end{bmatrix} = n - 1.$$

Using the fact that $\lambda \notin o(F_{11})$, we can perform some elementary row operations on $\begin{bmatrix} \lambda I_n - F \\ H \end{bmatrix}$ to get

$$\text{rank} \begin{bmatrix} \lambda I_n - F \\ H \end{bmatrix} = \text{rank} \begin{bmatrix} \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \mathbf{0}^T & 0 & f_{2,r} & \cdots & f_{2,r} & f_{2,r+1} \\ \mathbf{0}^T & 0 & 0 & \cdots & f_{3,r} & f_{3,r+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}^T & 0 & 0 & \cdots & 0 & f_{r,r+1} \\ \mathbf{0}^T & 0 & 0 & \cdots & 0 & 0 \\ 0 & \phi_1(\lambda) & \phi_2(\lambda) & \cdots & \phi_r(\lambda) & \phi_{r+1}(\lambda) \end{bmatrix}$$

where $\phi_j(\lambda) = \mathbf{h}_j + H_1(\lambda I_{n_j} - F_{11})^{-1} \mathbf{f}_{1j}$, $j = 2, \dots, r+1$ and \cdot denotes possible non-zero vectors. Since $f_{i,r+1} \neq 0$, $i = 2, \dots, r$, it follows that $\phi_2(\lambda) \neq 0$. Next, performing the

elementary row operations mentioned above on the matrix $\begin{bmatrix} \lambda I_n - F & G \\ H & 0 \end{bmatrix}$, we obtain

$$\text{rank} \begin{bmatrix} \lambda I_n - F & G \\ H & 0 \end{bmatrix} = n - r + \text{rank} \begin{bmatrix} f_{23} & f_{24} & & f_{2,r} & f_{2,r+1} \\ 0 & f_{34} & & f_{3,r} & f_{3,r+1} \\ & & & & \\ & & & & \\ 0 & 0 & & 0 & f_{r,r+1} \end{bmatrix}$$
$$\begin{bmatrix} \phi_1(\lambda) & \phi_2(\lambda) & \cdots & \phi_r(\lambda) & \phi_{r+1}(\lambda) & H_1(\lambda I_{n-r} - F_{11})^{-1} G_1 \end{bmatrix}$$

The right-hand side of the above equation is exactly the result that we would get if we were to apply the elementary

row operations to the matrix

$$\begin{bmatrix} \lambda I_{n-1} - \hat{F} & \hat{G} \\ \hat{H} & 0 \end{bmatrix}.$$

Therefore,

$$\text{rank} \begin{bmatrix} \lambda I_n - F & G \\ H & 0 \end{bmatrix} = \text{rank} \begin{bmatrix} \lambda I_{n-1} - \hat{F} & \hat{G} \\ \hat{H} & 0 \end{bmatrix}.$$

Since $f_{i,i+1} \neq 0, i = 2, \dots, r$, it follows from the structure of \hat{F} that $\lambda \notin \sigma(\hat{F})$. Therefore,

$$\text{rank} \begin{bmatrix} \lambda I_{n-1} - \hat{F} & \hat{G} \\ \hat{H} & 0 \end{bmatrix} = n - 1$$

if and only if λ is a blocking zero of $(\hat{F}, \hat{G}, \hat{H})$. The result of the theorem then follows, completing the proof.

Brief Paper

Qualitative Analysis and Decentralized Controller Synthesis for a Class of Large-scale Systems with Symmetrically Interconnected Subsystems*

M. K. SUNDARESHAN†‡ and R. M. ELBANNA†

Key Words—Decentralized control; large-scale systems; system analysis; control system synthesis; controllability and observability; Lyapunov methods.

Abstract—A number of large-scale interconnected systems often encountered in practice are composed of subsystems with similar dynamics interconnected in a symmetrical fashion and the synthesis of controllers for such systems must exploit the special structural properties in order to avoid overly conservative designs and to take advantage of the possible beneficial effects of the interconnections. An analysis of some important qualitative properties of such symmetrically interconnected systems focussing on the spectrum characterization, controllability and observability, and the solutions of the algebraic Riccati equation and the matrix Lyapunov equation is conducted in this paper and procedures for constructing the solutions to the analysis problems at the overall system level from the computationally simple subsystem level solutions are developed. A decentralized controller design procedure is presented as an illustration of the utilization of the available structural information in addressing synthesis problems. Numerical examples are included to demonstrate the superiority of the presented designs over the use of existing approaches which do not take full advantage of the structural knowledge in these large-scale systems.

1. Introduction

A FUNDAMENTAL structural feature often observed in several large-scale systems commonly encountered in the real world is the identical nature of the operational characteristics of the components, such as the subsystems whose interconnection constitutes the large system under consideration and the interconnection units used to interconnect these subsystems. Although true in several naturally evolving systems as well, this is particularly prevalent in engineering systems assembled by man, since a popular strategy for constructing a large system is to mass produce several smaller but identical components and to perform the assembly by interconnecting the subsystem blocks using the interconnection blocks. The reasons for employing such a strategy could be several; the principal ones being the use of automated and standardized design procedures for the mass production of components and the desire to employ massively parallel computation where a number of identical units are operated in parallel for

distributed processing to achieve a system-wide objective. The phenomenal growth of interest in neuro-computing and artificial neural networks in the recent times can be attributed to the latter reason (Hopfield, 1982; Rumelhart and McClelland, 1986). Such principles are employed in the synthesis of various kinds of systems ranging from microlevel implementations [such as in VLSI/VHSIC circuits (Mead, 1989)] to macrolevel assemblies [such as in power system networks comprising of identical generating units with balanced interconnections for exchange of generated power (Elgerd, 1981; Glavitsch and Gahuna, 1972)] and have also contributed to the evolution of several popular architectures [such as the star topology in local area computer networks (Stallings, 1984)].

This paper is concerned with the analysis of some fundamental properties of a class of dynamical large-scale systems assembled from interconnecting identical subsystems and with the identification of a few synthesis procedures that exploit the special structural arrangement present in these systems. Specifically, we shall consider a large-scale system \mathcal{S} that may be described as an interconnection of s subsystems $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_s$, by

$$\begin{aligned} \dot{x}_i(t) &= A_i x_i(t) + \sum_{j=1}^s H_{ij} x_j(t) + B_i u_i(t) \\ y_i(t) &= C_i x_i(t), \quad i = 1, 2, \dots, s, \end{aligned} \quad (1.1)$$

where $x_i(\cdot): R \rightarrow R^n$ and $y_i(\cdot): R \rightarrow R^p$. System (1.1) can be equivalently described by the composite equations:

$$\dot{x}(t) = Ax(t) + Bu(t); \quad y(t) = Cx(t) \quad (1.2)$$

where $x(\cdot): R \rightarrow R^{sn} \ni x^T(\cdot) = [x_1^T(\cdot) x_2^T(\cdot) \cdots x_s^T(\cdot)]$, $u(\cdot): R \rightarrow R^{sm} \ni u^T(\cdot) = [u_1^T(\cdot) u_2^T(\cdot) \cdots u_s^T(\cdot)]$ and $y(\cdot): R \rightarrow R^{sp} \ni y^T(\cdot) = [y_1^T(\cdot) y_2^T(\cdot) \cdots y_s^T(\cdot)]$, with the composite matrices $A \in R^{sn \times sn}$, $B \in R^{sm \times sn}$, and $C \in R^{sp \times sn}$ having the structure

$$A = \begin{bmatrix} A_1 & H_{12} & H_{13} & \cdots & H_{1s} \\ H_{12} & A_1 & H_{12} & \cdots & H_{12} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ H_{12} & H_{12} & H_{12} & \cdots & A_1 \end{bmatrix}$$

$B = \text{diag}[B_i]$, and $C = \text{diag}[C_i]$. We shall hereafter refer to this system as a Symmetrically Interconnected System.

It must be emphasized that one finds several modeling studies in very diverse areas resulting in state models of the type described by (1.1) and symmetrically interconnected systems arise as models for several physical processes. In fact, a number of these application areas are quite well-known to control specialists since these models were developed in the course of designing appropriate control strategies for these systems. For the sake of illustrating this

* Received 12 February 1990; revised 7 June 1990; received in final form 20 June 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor T. Başar under the direction of Editor A. P. Sage.

† Department of Electrical and Computer Engineering, University of Arizona, Tucson, Arizona 85721, U.S.A.

‡ Author to whom all correspondence should be addressed.

claim, we shall briefly mention a few specific examples of model developments (obtained especially for controller synthesis), giving references for details. A formulation of the load-frequency control problem in a multi-machine power system given in Mohadjer and Johnson (1983) and Abdulla (1986) results in a large scale interconnected system having the structure (1.1). Models having the present structure for industrial manipulators with several degrees of freedom are developed in Vukobratovic *et al.* (1977) and Vukobratovic and Stokic (1982) and are shown to be appropriate vehicles for the synthesis of improved control strategies. In Bryson and Ho (1969) can be found a model of the present type for a mechanical structure comprising of a number of spring-coupled pendulums for which control schemes using traditional optimal control theory are developed.

While the above examples serve to illustrate the occurrence of symmetrically interconnected system models in several applications where the models are developed for the explicit construction of control strategies, several other applications outside the mainstream control interests abound. For example, synthesis of neural networks employing the popular Hopfield-type models for optimization and input-output mapping applications are discussed in Pineda (1988) and Sudharsanan and Sundareshan (1989) where symmetrically interconnected system structures arise as network architectures of choice. In addition to their frequent occurrence in diverse applications mentioned above, symmetrically interconnected systems offer a very convenient framework for several investigations which could be used at the first step in the analysis and design studies of more general large scale systems and modifying the results at the later steps through appropriate perturbation approaches.

Symmetrically interconnected systems have been the subject of some study in the recent times, which have mainly focussed on the stability analysis of the composite system (Bergen, 1979; Baliga and Rao, 1980; Lunze, 1989). An important outcome resulting from these works is the conclusive demonstration that the knowledge of the structural information can be utilized for obtaining significantly improved results compared with the approaches that attempt to investigate the properties of general systems with arbitrary interconnections. It is evident that the latter approaches are generally more conservative, are of a weak-coupling nature (Jamshidi, 1983) and ignore the possible beneficial effects of the interconnection patterns thus disregarding the cooperative functioning of the subunits which underlies the synthesis of a majority of engineering systems. In particular, for several complex technological processes synthesized such that the subsystems must cooperate in order to execute a specific task, use of overly general approaches (i.e. results that are developed under very general settings in order to be valid for arbitrary system structures and system objectives) could indeed be disadvantageous and may simply not portray the reality of the environment.

In this paper, we shall examine some fundamental properties of symmetrically interconnected systems, mainly focusing on the eigenvalue characterization and the controllability/observability questions. These are further utilized in the development of solutions to the algebraic Riccati equation and the Lyapunov matrix equation which are encountered in several areas of system analysis and design. As specific illustrations of how the knowledge of the structural information can be exploited in the design of synthesis procedures, we shall present a few results for the synthesis of decentralized controllers for stabilization and pole placement. Through numerical examples, we shall demonstrate that these designs offer significant improvements over the existing approaches for this class of large-scale systems. The contributions of this paper are the various analysis results presented in Section 2 and two synthesis results presented in Section 3.

2. Analysis of qualitative properties

For the system \mathcal{S} described by (1.1) or (1.2), let us define

$$A_p \approx A_1 + (s-1)H_{12} \begin{pmatrix} I_{12} \end{pmatrix} \quad (2.1)$$

We then have the following results.

Lemma 1 (Eigenvalue characterization).

$$\text{spec}(A) = \text{spec}(A_p) \cup \left\{ \bigcup_{i=1}^s \text{spec}(A_m) \right\} \quad (2.2)$$

where $\bigcup_{i=1}^s \text{spec}(A_m) = \text{spec}(A_m) \cup \text{spec}(A_m) \cup \dots \cup \text{spec}(A_m)$ (i.e. eigenvalues of A_m repeated $(s-1)$ times).

Proof. Consider the matrix $T \in R^{(n+m)s}$ given by

$$T = \begin{pmatrix} I & 0 & 0 & \dots & 0 \\ -I & I & 0 & \dots & 0 \\ -I & 0 & I & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -I & 0 & 0 & \dots & 0 \end{pmatrix} \quad (2.3)$$

where I is an $n \times n$ identity matrix. Then,

$$\tilde{A} = TAT^{-1} = \begin{pmatrix} A_p & H_{12} & H_{12} & \dots & H_{12} \\ 0 & A_m & 0 & \dots & 0 \end{pmatrix} \quad (2.4)$$

Hence the result follows. \blacksquare

Lemma 1 is useful in analyzing the stability of the uncontrolled system i.e. system \mathcal{S} with $u=0$. It may be noted that evaluation of the eigenvalues of the smaller dimensional matrices $A_p \in R^{n \times n}$ and $A_m \in R^{m \times m}$ is sufficient to conclude the stability of the overall system \mathcal{S} of dimension $sn \times sn$.

Lemma 2 (Complete controllability). System \mathcal{S} is completely controllable (c.c.) if and only if the pairs (A_p, B_1) and (A_m, B_1) are both c.c.

Proof. Using the transformation matrix T given in (2.3), $\tilde{A} = TAT^{-1}$ is obtained as in (2.4) and further,

$$\tilde{B} = TB = \begin{pmatrix} B_1 & 0 & 0 & \dots & 0 \\ -B_1 & B_1 & 0 & \dots & 0 \\ -B_1 & 0 & B_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -B_1 & 0 & 0 & \dots & 0 \end{pmatrix} \quad (2.5)$$

Now consider

$$M = [\alpha I - \tilde{A} : \tilde{B}] = \begin{pmatrix} \alpha I - A_p & -H_{12} & -H_{12} & B_1 & 0 & 0 & \vdots \\ 0 & \alpha I - A_m & 0 & 0 & 0 & 0 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & \alpha I - A_m & -B_1 & 0 & 0 & \vdots \end{pmatrix}$$

Then, it follows (Chen, 1984) that

$$(A, B) \text{ c.c.} \Leftrightarrow (\tilde{A}, \tilde{B}) \text{ c.c.} \Leftrightarrow \text{Rank}(M) = sn. \quad (2.6)$$

Let $D \in R^{(n+m)s \times (n+m)s}$ be constructed in the form

$$D = \begin{pmatrix} I & 0 & 0 & \dots & 0 \\ 0 & I & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & I \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} I & 0 & 0 & \dots & 0 \\ I & I & 0 & \dots & 0 \\ 0 & I & 0 & I & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ I & 0 & 0 & \dots & I \end{pmatrix}$$

Then,

$$\tilde{M} = MD =$$

$$\begin{bmatrix} \alpha I - A_p & -H_{12} & \cdots & -H_{12} & B_1 & 0 & \cdots & 0 \\ 0 & \alpha I - A_m & \cdots & 0 & 0 & B_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha I - A_m & 0 & 0 & \cdots & B \end{bmatrix}$$

and hence,

$$\text{Rank}(\tilde{M}) = \text{Rank}(\tilde{M}) = \text{Rank}[\alpha I - A_p : B_1] + (s-1) \cdot \text{Rank}[\alpha I - A_m : B_1]. \quad (2.7)$$

Since $\text{Rank}[\alpha I - A_p : B_1]$ and $\text{Rank}[\alpha I - A_m : B_1]$ cannot exceed n , it follows from (2.4), (2.5) and (2.6) that (A, B) c.c. $\Leftrightarrow (A_p, B_1)$ c.c., and (A_m, B_1) c.c. On the other hand, if (A, B) is not c.c., then $\text{Rank}(\tilde{M}) < sn$, which implies that either or both of (A_p, B_1) and (A_m, B_1) are not c.c. ■

Sufficient conditions for the complete controllability of both (A_p, B_1) and (A_m, B_1) can be obtained as

- (i) (A_1, B_1) is c.c.
- (ii) $\text{Rank}[B_1 : H_{12}] = \text{Rank}[B_1]$.

For details, one may refer to Elbanna (1988).

By duality, one can obtain corresponding results for complete observability, which we shall state without proof in the following.

Lemma 3 (Complete observability). System \mathcal{F} is c.o. if and only if the pairs (A_p, C_1) and (A_m, C_1) are both c.o.

Of fundamental interest in the theory of linear dynamical systems are the solutions to the algebraic Riccati equation and the Lyapunov matrix equation, which are needed for conducting various types of designs of controllers and estimators. For the symmetrically interconnected system under consideration, such solutions can be constructed simply from solving corresponding equations for considerably lower order systems. The following results provide the details of these constructions. We shall assume that (A, B) is c.c.

Theorem 1. Let $P_p \in R^{n \times n} \ni P_p = P_p^T > 0$ and $P_m \in R^{n \times n} \ni P_m = P_m^T > 0$ denote the solutions of the Riccati equations

$$A_p^T P_p + P_p A_p - P_p B_1 R_1^{-1} B_1^T P_p + Q_1 = 0 \quad (2.8)$$

$$A_m^T P_m + P_m A_m - P_m B_1 R_1^{-1} B_1^T P_m + Q_1 = 0 \quad (2.9)$$

respectively, for arbitrarily selected $Q_1 \in R^{n \times n} \ni Q_1 = Q_1^T > 0$ and $R_1 \in R^{m \times m} \ni R_1 = R_1^T > 0$, where A_p and A_m are given by (2.1). Then the unique symmetric and positive definite solution $P \in R^{sn \times sn}$ of the Riccati equation for the composite system (1.2), i.e.

$$A^T P + PA - PBR^{-1}B^T P + Q = 0 \quad (2.10)$$

where $Q = \text{diag}[Q_1]$ and $R = \text{diag}[R_1]$, has the structure

$$P = \begin{bmatrix} P_1 & P_2 & \cdots & P_2^T \\ P_2 & P_1 & \cdots & P_2^T \\ \vdots & \vdots & \ddots & \vdots \\ P_2^T & P_2 & \cdots & P_1 \end{bmatrix} \quad (2.11)$$

where $P_1 \in R^{n \times n}$ and $P_2 \in R^{n \times n}$ are given by

$$\begin{aligned} P_1 &= \{P_p + (s-1)P_m\}/s \\ \text{and } P_2 &= \{P_p - P_m\}/s \end{aligned} \quad (2.12)$$

Proof. For the selected Q and R matrices, the solution P of (2.10) is unique, symmetric and positive definite, since the pair (A, B) is c.c. Let the $n \times n$ matrix partitions of this solution be given by

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1s} \\ P_{21} & P_{22} & \cdots & P_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ P_{s1} & P_{s2} & \cdots & P_{ss} \end{bmatrix} \quad (2.13)$$

Our objective is to demonstrate that $P_{ii} = P_1 \forall i = 1, 2, \dots, s$ and $P_{ij} = P_2 \forall i, j = 1, 2, \dots, s, i \neq j$.

With P partitioned as in (2.13), expanding (2.10) one obtains the two sets of equations,

$$A_i^T P_{ii} + P_{ii} A_i - P_{ii} S_i P_{ii} + \hat{Q}_i = 0, \quad i = 1, 2, \dots, s, \quad (2.14)$$

and

$$\begin{aligned} A_i^T P_{ij} + P_{ij} A_j + P_{ij} (H_{12} - S_i P_{ii}^T) + (H_{12} - S_i P_{ii}^T)^T P_{ij} \\ + \sum_{k=1}^s \left[P_{ik} \left(H_{12} - \frac{1}{2} S_i P_{ik}^T \right) + \left(H_{12} - \frac{1}{2} S_i P_{ik}^T \right)^T P_{ik} \right] = 0, \end{aligned} \quad (2.15)$$

where $S_i = B_1 R_1^{-1} B_1^T$, $\hat{Q}_i = Q_1 + \sum_{j=1, j \neq i}^s (G_{ij} + G_{ij}^T)$, $G_{ij} = P_{ij} [H_{12} - \frac{1}{2} S_i P_{ij}^T]$. Now, letting $P_{ij} = P_j \forall i, j = 1, \dots, s, i \neq j$, one can see that $G_{ij} + G_{ij}^T = H_{12}^T P_j + P_j H_{12} - P_j S_i P_j$ and $\hat{Q}_i = Q_1 + (s-1)[H_{12}^T P_i + P_i H_{12} - P_i S_i P_i]$. It hence follows that the solutions P_{ii} of (2.14) are equal for all i . Let $P_1 = P_{ii}$ denote this solution. With these substitutions, (2.14) and (2.15) can be rewritten as

$$\begin{aligned} A_1^T P_1 + P_1 A_1 - P_1 S_1 P_1 + Q_1 + (s-1) \\ \times [H_{12}^T P_2 + P_2 H_{12} - P_2 S_1 P_2] = 0 \end{aligned} \quad (2.16)$$

and

$$\begin{aligned} A_i^T P_2 + P_2 A_i + P_i (H_{12} - S_i P_2) + (H_{12} - P_2 S_i) P_i \\ + (s-2)[P_2 (H_{12} - \frac{1}{2} S_i P_2) + (H_{12} - \frac{1}{2} S_i P_2) P_2] = 0. \end{aligned} \quad (2.17)$$

Now, multiplying (2.17) by $(s-1)$ and adding the result to (2.16), one obtains

$$A_p^T P_p + P_p A_p - P_p S_1 P_p + Q_1 = 0 \quad (2.18)$$

where

$$P_p = P_1 + (s-1)P_2 \quad (2.19)$$

Similarly, subtracting (2.17) from (2.16), one obtains

$$A_m^T P_m + P_m A_m - P_m S_1 P_m + Q_1 = 0 \quad (2.20)$$

where

$$P_m = P_1 - P_2. \quad (2.21)$$

Since (A, B) is c.c., (A_p, B_1) and (A_m, B_1) are both c.c. from Lemma 2, which further ensures that the solutions P_p and P_m of (2.18) and (2.20) respectively are unique, symmetric and positive definite.

Solving (2.19) and (2.21) for P_1 and P_2 in terms of P_p and P_m results in the expressions given in (2.12). Also P in (2.13) reduces to the structure given in (2.11). Furthermore, using the transformation matrix T given in (2.3), one obtains

$$\begin{aligned} \hat{P} = TPT^{-1} = \begin{bmatrix} P_p & P_2 & P_2 & \cdots & P_2 \\ 0 & P_m & 0 & \cdots & 0 \\ 0 & 0 & P_m & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & P_m \end{bmatrix} \end{aligned}$$

and hence

$$\text{Spec}(P) = \text{Spec}(\hat{P}) = \text{Spec}(P_p) \cup \left\{ \bigcup \text{Spec}(P_m) \right\}.$$

Since P_p and P_m are positive definite, this concludes that P given by (2.11) is symmetric and positive definite. Finally, since the positive definite solution of (2.10) is unique, it follows that P given by (2.11) is the only solution of (2.10) and hence, our proof is complete. ■

The remarkable simplification resulting from the need to solve Riccati equations of considerably low dimensions only is evident from the above result. This property is indeed fundamental in our being able to synthesize efficient decentralized controllers and estimators for this class of systems. By employing a similar construction, the solution of the Lyapunov matrix equation for symmetrically interconnected systems can be obtained from solving corresponding equations of a much lower order, as given in the following result.

Theorem 2. Let $\text{Spec}(A_p) \in LHP$ and $\text{Spec}(A_m) \in LHP$, and let $P_p \in R^{n \times n} \ni P_p + P_p^T > 0$ and $P_m \in R^{m \times m} \ni P_m + P_m^T > 0$ denote the solutions of the Lyapunov matrix equations

$$\begin{aligned} A_p^T P_p + P_p A_p + Q_1 &= 0 \\ A_m^T P_m + P_m A_m + Q_1 &= 0 \end{aligned} \tag{2.22}$$

respectively, for an arbitrarily selected $Q_1 \in R^{(n+m) \times (n+m)} \ni Q_1 = Q_1^T > 0$, where A_p and A_m are given by (2.1). Then the unique symmetric and positive definite solution $P \in R^{(n+m) \times (n+m)}$ of the Lyapunov matrix equation for the composite system (1.2), i.e.

$$A^T P + P A + Q = 0 \tag{2.23}$$

where $Q = \text{diag}[Q_1]$, has the structure

$$\begin{bmatrix} P_1 & P_2 & P_3 \\ P_1^T & P_2^T & P_3^T \\ P_3 & P_2 & P_1 \end{bmatrix}$$

where $P_1 \in R^{n \times n}$ and $P_2 \in R^{n \times m}$ are given by

$$P_1 = \frac{1}{s} [P_p + (s-1)P_m] \tag{2.24}$$

$$\text{and } P_2 = \frac{1}{s} [P_p - P_m]$$

Proof. The proof follows along identical lines to that of Theorem 1 and noting that $\text{Spec}(A_p) \in LHP$ and $\text{Spec}(A_m) \in LHP$ together imply from Lemma 1, $\text{Spec}(A) \in LHP$. Hence, (2.23) has a unique symmetric and positive definite solution ■

3. Synthesis of controller algorithms

By making essential use of the properties of symmetrically interconnected systems outlined in the last section, it is possible to give systematic synthesis procedures for various types of controllers. For the sake of brevity, we shall give the details of only a few illustrative constructions. A greater variety of controller synthesis procedures and a detailed evaluation of the properties of these controllers can be found in Elbanna (1988). The specific design procedures that will be discussed here are based on the following results.

Theorem 3. For system \mathcal{J} described by (1.1) or (1.2), let $K_1 \in R^{m \times n} \ni \text{Spec}(A_1 - B_1 K_1) \in LHP$ and let $W \in R^{n \times n}$ given by

$$W = I + (\tau - 1)(H_{12}^T N_1 + N_1 H_{12}) \tag{3.1}$$

is positive definite for $\tau = s$ and $\tau = 0$, where $N_1 \in R^{n \times n} \ni N_1 = N_1^T > 0$ is the solution of

$$(A_1 - B_1 K_1)^T N_1 + N_1 (A_1 - B_1 K_1) + I = 0, \tag{3.2}$$

I being the $n \times n$ identity matrix. Then $\text{Spec}(A_p - B_1 K_1) \in LHP$ and $\text{Spec}(A_m - B_1 K_1) \in LHP$, where A_p and A_m are given by (2.1). Furthermore, $\text{Spec}(A - BK) \in LHP$, where $K \in R^{m \times (n+m)} \ni K = \text{diag}[K_1]$.

Proof. Starting with the system

$$\dot{z}(t) = (A_p - B_1 K_1)z(t) \tag{3.3}$$

where $z(\cdot): R \rightarrow R^n$ and computing the time-derivative of

$$V(z) = z^T(t)N_1 z(t) \tag{3.4}$$

along the trajectories of (3.3), one obtains

$$\begin{aligned} \dot{V}(z) &= z^T(t)[(A_p - B_1 K_1)^T N_1 + N_1 (A_p - B_1 K_1)]z(t) \\ &= z^T(t)[(A_1 - B_1 K_1)^T N_1 + N_1 (A_1 - B_1 K_1) \\ &\quad + (s-1)(H_{12}^T N_1 + N_1 H_{12})]z(t) \\ &= z^T(t)[-I + (s-1)(H_{12}^T N_1 + N_1 H_{12})]z(t). \end{aligned} \tag{3.5}$$

Hence, $\dot{V}(z)$ is negative definite if W in (3.1) is positive definite for $\tau = s$, which implies $\text{Spec}(A_p - B_1 K_1) \in LHP$. Similarly, it can be shown that $\text{Spec}(A_m - B_1 K_1) \in LHP$, if W in (3.1) is positive definite for $\tau = 0$.

Furthermore, following the steps in the proof of Lemma 1 (i.e. using the transformation matrix T given in (2.3)), it can be shown that

$$\text{Spec}(A - BK) = \text{Spec}(A_p - B_1 K_1) \cup \bigcup \text{Spec}(A_m - B_1 K_1)$$

and hence

$$\text{Spec}(A - BK) \in LHP.$$

The usefulness of the above result in the construction of informationally decentralized stabilizing controllers stems from the property that $K = \text{diag}[K_1]$. For the selection of K_1 given the pair (A_1, B_1) , one may be guided by the following considerations. If (A_1, B_1) is c.c., there exist in general an infinite number of selections of K_1 that will make $\text{Spec}(A_1 - B_1 K_1) \in LHP$. We are however interested in the choice of K_1 that will possibly make W positive definite for both $\tau = 0$ and $\tau = s$. Observing that a sufficient condition for W to be positive definite for both $\tau = 0$ and $\tau = s$ is

$$\|H_{12}\|^{-1} < \frac{1}{2(s-1)\lambda_M(N_1)} \tag{3.6}$$

where $\lambda_M(N_1)$ denotes the maximum eigenvalue of N_1 , the preferred choice of K_1 is the one that will result in a solution N_1 of the Lyapunov matrix equation (3.2) with the lowest value for $\lambda_M(N_1)$. An analytical solution to this problem is highly complicated (especially when the dimension of A_1 is large) since one is required to solve $n(n-2)$ nonlinear inequalities. Available literature on the Lyapunov matrix equation (Karanam, 1981; Mori *et al.*, 1986) gives only the bounds on $\lambda_M(N_1)$, but these cannot be used for the problem at hand. Approximate numerical solutions can however be determined for each dimension of $(A_1 - B_1 K_1)$ when the pair (A_1, B_1) is in certain canonical forms, and iteratively changing its eigenvalues. For the sake of illustration, the results of this computation for $n = 2, 3, 4$ and 5 for the eigenvalue locations to yield the smallest value of $\lambda_M(N_1)$ in the case when (A_1, B_1) is in the controllable canonical form are given in Table 1.

When (3.1) is not satisfied by the interconnection pattern existing in the system to permit the construction of a diagonal K matrix according to Theorem 3, an alternate construction of K can be made for assigning the eigenvalues of $(A - BK)$ at arbitrarily specified locations, as given in the following result.

Theorem 4. For system \mathcal{J} described by (1.1) (or (1.2)), let $K_i \in R^{m \times n}$, $i = 1, 2, \dots, s$, exist such that

$$\text{spec}(A_p - B_1 K_1) = \Gamma_1 \tag{3.7}$$

and

$$\text{spec}(A_m - B_1 K_1) = \Gamma_j, \quad j = 2, 3,$$

where A_p and A_m are given by (2.1) and Γ_i , $i = 1, 2, \dots, s$ are sets of arbitrarily specified complex numbers $\Gamma_i = \{\lambda_1^i, \lambda_2^i, \dots, \lambda_n^i\}$ satisfying the condition that for any $\lambda_j^i \in \Gamma_i$, $\lambda_j^i \notin R \ni \lambda_j^{i*} \in \Gamma_i(\lambda_j^i)^*$ denotes the conjugate of λ_j^i . Then

$$\text{spec}(A - BK) = \bigcup \Gamma_i \tag{3.8}$$

TABLE 1

n	Eigenvalue locations for minimum $\lambda_M(N_1)$	$\lambda_M(N_1)$
2	$-1 \pm 1.4142j$	1.3605
3	$-1, -1.2679, -4.7321$	3.024
4	$-0.5 \pm 0.866j, -1.0436, -23.9564$	8.177
5	$-0.3 \pm 0.714j, -1.383, -3.618, -5$	24.0635

for $K \in R^{sm \times sm}$ given by

$$\begin{aligned} & K_1 \quad 0 \quad 0 \quad \cdots \quad 0 \\ & K_1 - K_2 \quad K_2 \quad 0 \quad \cdots \quad 0 \\ & K_1 - K_3 \quad 0 \quad K_3 \quad \cdots \quad 0 \\ & \vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \\ & K_1 - K_s \quad 0 \quad 0 \quad \cdots \quad K_s \end{aligned} \quad (3.9)$$

Proof. The proof follows simply by observing that $(A - BK)$ can be transformed using the transformation matrix T given by (2.3) into

$$D = T(A - BK)T^{-1} = \begin{bmatrix} A_p - B_1 K_1 & H_{12} & H_{12} & H_{12} \\ 0 & A_m - B_1 K_2 & 0 & 0 \\ 0 & 0 & A_m - B_1 K_3 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & A_m - B_1 K_s \end{bmatrix}$$

and hence, $\text{spec}(A - BK) = \text{spec}(D) = \text{spec}(A_p - B_1 K_1) \cup \text{spec}(A_m - B_1 K_2) \cup \text{spec}(A_m - B_1 K_3) \cup \cdots \cup \text{spec}(A_m - B_1 K_s)$. ■

It should be noted that K in (3.9), although not diagonal, has a special structure and hence does not correspond to the gain matrix of a centralized controller; in fact, it results in a decentralized control scheme with a minimal information exchange in that subsystem \mathcal{S}_j , $j \neq 1$, receives an additional input $(K_1 - K_j)x_1$. If the design objective is not arbitrary eigenvalue assignment, but stabilization of the overall system, Γ_1 may be identically chosen for all $j = 2, 3, \dots, s$, which implies $K_j = K_1, \dots, K_s$. Furthermore, if $K_1 \in R^{m \times m}$ exists such that $\text{spec}(A_p - B_1 K_1) \subset LHP$ and $\text{spec}(A_m - B_1 K_1) \subset LHP$, then the choice of $K_j = K_1 \forall j = 1, 2, 3, \dots, s$, results in a diagonal K and hence a corresponding decentralized control scheme $u = -Kx$.

A particularly noteworthy feature of the present designs are their simplicity in the construction of controller gains compared to similar controller design procedures for large-scale systems one may find in the literature. There are no iterative calculations that are needed in the determination of controller gains, which are mainly computed by simple pole placement computations. The present designs also follow precisely executed steps in the sense that there are no trial and error computations needed as in several available designs for these types of systems.[†] In addition to the computational simplicity and precision, the present design also permits elements of considerably larger magnitudes in the interconnection matrix. This is of considerable interest since most of the available results for large-scale systems are of a weak coupling nature and impose very restrictive conditions on the norm of the interconnection matrix for the design to be valid. Some numerical examples to illustrate this claim will be given in the next section. It is to be emphasized that these advantages are a result of our being able to exploit the available structural knowledge of the interconnections present in the overall system.

By employing duality arguments, corresponding designs for constructing observers for symmetrically interconnected systems can be developed, which possess several advantages over the decentralized observer design procedures available in the literature (Sundareshan, 1977; Sundareshan and Huang, 1984). For a concise description of the procedure,

[†] The constructions are often based on an initial choice of Lyapunov functions for the isolated subsystems and the use of stability arguments for the overall system following either a vector Lyapunov function approach or a weighted sum approach (Jamshidi, 1983). The trial and error nature of the design stems from the fact that if the interconnections do not satisfy the required conditions, one is obliged to start with an alternate selection of the subsystem Lyapunov functions and repeat the entire process.

consider the observation scheme

$$\begin{aligned} \hat{x}_1(t) &= (A_1 - K_1 C_1) \hat{x}_1(t) + K_1 y_1(t) + B_1 u_1(t) \\ &+ \sum_{i=2}^s \{ [H_{12} - (K_1 - K_i) C_i] \hat{x}_i(t) \\ &+ (K_1 - K_i) y_i(t) \} \end{aligned} \quad (3.10)$$

and

$$\begin{aligned} \hat{x}_i(t) &= (A_i - K_i C_i) \hat{x}_i(t) + K_i y_i(t) + B_i u_i(t) \\ &+ \sum_{j=1}^s H_{ij} \hat{x}_j(t), \quad i = 2, 3, \dots, s, \end{aligned} \quad (3.11)$$

which together with (1.1) results in the "error system"

$$\dot{e}_1(t) = (A_1 - K_1 C_1) e_1(t) + \sum_{i=2}^s [H_{12} - (K_1 - K_i) C_i] e_i(t) \quad (3.12)$$

and

$$\dot{e}_i(t) = (A_i - K_i C_i) e_i(t) + \sum_{j=1}^s H_{ij} e_j(t), \quad i = 2, 3, \dots, s, \quad (3.13)$$

where $e_i(t) = x_i(t) - \hat{x}_i(t)$, $i = 1, 2, \dots, s$. Observing that the problem of interest now is the selection of the observer gain matrices $K_i \in R^{n \times n}$, $i = 1, 2, \dots, s$, such that (3.12) and (3.13) are asymptotically stable, one may readily construct, by duality, observer design algorithms that parallel the controller design algorithms given above. More details on this construction, together with additional discussion on the properties of the observer, may be found in Elbanna (1988).

Some examples:

Example 1. Consider the Symmetrically Interconnected System

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + \sum_{j=1}^s H_{ij} x_j(t);$$

$$y_i(t) = C_i x_i(t), \quad i = 1, 2,$$

where

$$H_{12} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 3.254 & -1.472 & 4.375 \\ 0.0718 & 0.828 & 0.2584 \\ 0.003 & 0.2752 & 0.3084 \\ 0.1538 & 0.2746 & 0.1642 \end{bmatrix}.$$

$B_1 = [0 \ 0 \ 1]^T$ and C_1 is arbitrary.

Selecting $K_1 = [9.2542 \ 10.528 \ 11.375]$ to place the eigenvalues of $(A_1 - B_1 K_1)$ at -1 , -1.2679 and -4.7321 (see table for eigenvalue locations given earlier) and solving (3.2) for N_1 , one obtains

$$N_1 = \begin{bmatrix} 1.6944 & -0.501 & -0.6667 \\ -0.501 & 0.6667 & -0.501 \\ -0.6667 & -0.501 & 1.502 \end{bmatrix}$$

Now constructing W_1 and W_2 as

$$W_1 = I + (H_{12}^T N_1 + N_1 H_{12}) = \begin{bmatrix} 1.9242 & -0.0488 & 0.0604 \\ -0.0488 & 0.9445 & -0.3471 \\ 0.0604 & -0.3471 & 0.9871 \end{bmatrix}$$

and

$$W_2 = I - (H_{12}^T N_1 + N_1 H_{12}) = \begin{bmatrix} 0.0708 & 0.0488 & -0.0604 \\ 0.0488 & 1.0555 & 0.3471 \\ -0.0604 & 0.3471 & 1.02129 \end{bmatrix}.$$

one can check that W_1 and W_2 are positive definite. Hence, the decentralized control scheme $u_i(\cdot) = -K_i x_i(\cdot)$, $i = 1, 2$, stabilizes the given system.

For comparison with available methods, it may be noted that the interconnection matrices in this system do not satisfy the required conditions for decentralized stabilization by any of the earlier developed results, including the most general and recently developed results due to Ikeda *et al.* (1983) and Shi and Gao (1986). This example conclusively demonstrates

the superiority of design procedures such as the present one which constructively utilize the available knowledge of the interconnection structure within the system when compared with design procedures that attempt to stabilize a broad class of large-scale systems with arbitrary interconnection structures, which naturally yield overly conservative (and hence less useful in practice) results

Example 2 Consider the Symmetrically Interconnected System

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + \sum_{j=1}^3 H_{ij} x_j(t),$$

$$y_i(t) = C_i x_i(t), \quad i = 1, 2, 3,$$

where

$$A_i = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 0 & 1 \\ 1 & 1 & 0.5 \end{bmatrix}$$

$$H_{ij} = \begin{bmatrix} 0.527 & 0.6792 & 0.1278 \\ -0.095 & -0.1633 & -0.2729 \\ 0.4532 & 0.3335 & 0.606 \end{bmatrix}$$

$B_i = [1 \ -2 \ 4]^T$ and C_i is arbitrary.

Following the steps of the design algorithm, one can determine that a selection of $K_i \in R^{1 \times 3}$ such that W_i and W_c are both positive definite cannot be made. Hence we follow the construction in Step 4. Since $A_p = A_1 + 2H_{11}$ and $A_m = A_1 + H_{11}$ are

$$A_p = \begin{bmatrix} 2.054 & 1.3584 & 3.2556 \\ -0.19 & -0.3266 & 0.4542 \\ 1.9064 & 1.667 & 0.712 \end{bmatrix}$$

and

$$A_m = \begin{bmatrix} 0.473 & -0.6792 & 2.8722 \\ 0.095 & 0.1633 & 1.2729 \\ 0.5468 & 0.6665 & -0.106 \end{bmatrix}$$

and the pairs (A_p, B_1) and (A_m, B_1) are c.e., one may select K_i , $i = 1, 2, 3$, as $K_1 = [1.4638 \ 1.3224 \ 1.7801]$, $K_2 = [1.5847 \ 10.9368 \ 28.9674]$, $K_3 = [1.7535 \ 28.4674 \ 17.4279]$, to make $\text{spec}(A_p - B_1 K_1) = \{-2.5, -0.5 \pm j0.5\}$, $\text{spec}(A_m - B_1 K_1) = \{-3, -3.5, -4\}$ and $\text{spec}(A_m - B_1 K_1) = \{-4.5, -5, -5.5\}$ (these eigenvalue locations were selected arbitrarily for illustrative purposes). The control scheme $u = -Kx$, where K has the structure given in (3.9), results in a composite system with eigenvalues at the specified locations i.e. $\{-0.5 \pm j0.5, -2.5, -3, -3.5, -4, -4.5, -5, -5.5\}$.

It should be emphasized that no existing result for interconnected systems performs the pole placement in such a simple manner.

4. Conclusions

The principal contributions of this paper are the analysis of the qualitative properties of an important class of large-scale interconnected systems composed of symmetrically interconnected subsystems and a decentralized controller synthesis procedure which makes essential use of these properties. While a number of specific important results are developed in this paper, they should be considered as illustrations of a more fundamental mechanism by which solutions to the analysis and synthesis problems at the overall system level can be developed from subsystem level solutions. In particular, a demonstration is given of how the synthesis of controllers can exploit the special structural properties of the systems in order to avoid overly conservative designs and also to take advantage of the possible beneficial effects of the interconnection patterns present. Evidently, such a strategy of tailoring the design procedure to the existing interconnection patterns has several payoffs, some of which are quantitatively illustrated through numerical examples in this paper.

Although the primary focus in this paper is to develop the analysis and synthesis results that make use of the specified structural arrangement of the subsystems, some extended results can be developed through standard perturbation approaches to cover the cases when the interconnection

pattern departs from the nominally specified structure or when the subsystem dynamics are altered due to parameter variations. An illustration of such an approach, for the particular problem of stability analysis, can be found in Lunze (1989).

References

- Abdulla, A. M. (1986). New techniques for the load-frequency control of multi-area power systems. Ph.D. Dissertation, University of Arizona, Tucson, AZ.
- Baliga, G. V. and M. V. C. Rao (1980). On symmetric and unity interconnections between three nonlinear subsystems. *Automatica*, **25**, 561-570.
- Bergen, A. R. (1979). Results on complementary coupling of similar systems. *Proc. Int. Symp. on Circuits and Systems*, Tokyo, 412-416.
- Bryson, A. E. and Y. C. Ho (1969). *Applied Optimal Control*, Blaisdell, Waltham.
- Chen, C. T. (1984). *Linear Systems: Theory and Design*. Holt, Reinhart and Winston, New York.
- Elbanna, R. (1988). Some new results on the stabilization and state estimation in large-scale systems. Ph.D. Dissertation, University of Arizona, Tucson, AZ.
- Elgerd, O. I. (1981). *Electric Energy System Theory: An Introduction*, McGraw-Hill, New York.
- Glavitsch, H. and F. D. Galiana (1972). Load-frequency control with particular emphasis on thermal power stations. In Handschin, E. (Ed.), *Real-time Control of Electric Power Systems*. Elsevier, Amsterdam.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 2554-2558.
- Ikeda, M., D. D. Siljak and K. Yasuda (1983). Optimality of decentralized control for large-scale systems. *Automatica*, **19**, 309-316.
- Jamshidi, M. (1983). *Large-scale Systems: Modeling and Control*, North-Holland, New York.
- Karanam, V. R. (1981). Lower bounds on the solutions of Lyapunov matrix and algebraic Riccati equations. *IEEE Trans. Aut. Control*, **AC-26**, 1288-1290.
- Lunze, J. (1989). Stability analysis of large-scale systems composed of strongly connected similar subsystems. *Automatica*, **25**, 561-570.
- Mead, C. E. (1989). *Analog VLSI and Neural Systems*. Addison-Wesley, New York.
- Mohadjer, M. and C. D. Johnson (1983). Power system control with disturbance accommodation. *Proc. 22nd IEEE Conf. on Decision and Control*, 1429-1433.
- Mori, T., N. Fujuma and M. Kiawahar (1986). Explicit solution and eigenvalue bounds in the Lyapunov matrix equations. *IEEE Trans. Aut. Control*, **AC-31**, 656-658.
- Pineda, F. J. (1988). Dynamics and architecture for neural computation. *J. Complexity*, **4**, 216-245.
- Rumelhart, D. E. and J. L. McClelland (1986). *Parallel Distributed Processing*, MIT Press, Cambridge, MA.
- Shi, Z. and W. B. Gao (1986). Stabilization by decentralized control for large-scale interconnected systems. *Large-scale Syst.*, **10**, 147-155.
- Sudharsanan, S. I. and M. K. Sundareshan (1989). Neural network computational algorithms for least-squares estimation problems. *Proc. 1989 Int. Joint Conf. on Neural Networks (IJCNN '89)*, Washington, D.C.
- Sundareshan, M. K. (1977). Decentralized observation in large-scale systems. *IEEE Trans. Syst. Man, Cybern.*, **SMC-7**, 863-868.
- Sundareshan, M. K. and P. C. K. Huang (1984). On the design of a decentralized observation scheme for large-scale systems. *IEEE Trans. Aut. Control*, **AC-29**, 274-276.
- Stallings, W. (1984). *Local Networks: An Introduction*. Macmillan, New York.
- Vukobratovic, M. and D. M. Stokic (1982). *Control of Manipulation Robots: Theory and Application*. Springer, Berlin.
- Vukobratovic, M., D. M. Stokic and D. S. Hristic (1977). New control concept of anthropomorphic manipulators in industrial applications. *IFTOMM J. Mechan. Machine Theory*, **12**, 100-103.

On Practical Stability of Linear Multivariable Feedback Systems With Time-delays*

WEINING FENG†

Key Words—Linear multivariable feedback; infinitesimal perturbations; practical stability; time-delays; Smith Predictor.

Abstract—In this paper, practical stability properties of linear multivariable feedback systems with time-delays are studied. The control schemes considered are conventional feedback control and Smith Predictor control. Depending upon the known perturbation structures, tight conditions are given which guarantee practical stability of the control system.

Nomenclature

\mathcal{R}	Field of real numbers.
\mathbb{C}	Field of complex numbers.
$\mathcal{R}^{m \times n}$	Matrices with m rows and n columns with elements in \mathcal{R} .
$\mathbb{C}^{m \times n}$	Matrices with m rows and n columns with elements in \mathbb{C} .
$\mathcal{R}(s)^{m \times n}$	Matrices with m rows and n columns with elements being rational transfer functions.
$\alpha \in A$	Element α belongs to set A .
$ c $	Absolute value of $c \in \mathbb{C}$.
M^+	Matrix M with elements being replaced by their absolute values.
$\text{Re}(c)$	The real part of $c \in \mathbb{C}$.
$\lambda_i(M)$	i th eigenvalue of the matrix $M \in \mathbb{C}^{m \times n}$.
$\sigma_{\max}(M)$	Largest singular value of matrix $M \in \mathbb{C}^{m \times n}$.
$\rho(M)$	Spectral radius, $\max \lambda_i(M) $; $i = 1, 2, \dots, n$; $M \in \mathbb{C}^{n \times n}$.
$\mu(M)$	Structured singular value of matrix $M \in \mathbb{C}^{m \times n}$.
d	A generic symbol for some (small but not infinitesimal) negative number.
s_*	A generic symbol for some complex number, $s_* = \sigma + j\omega$ with $\sigma \geq d$ and $ s_* \rightarrow +\infty$.

1. Introduction

ONE OF the difficulties with time-delay system analysis is that the poles of the system are infinite in number. There are generally two ways of treating delay elements to eliminate such terms from the characteristic equations of closed-loop systems. For systems with relatively small delay constants, the time-delays are ignored or replaced by sufficiently accurate Padé approximations (Takahashi *et al.*, 1987). The conventional structure of feedback control can then be applied as shown by Fig. 1. If delay constants are considerably larger but the system is open-loop stable, a special time-delay compensator such as the Smith Predictor (Fig. 2) is usually a candidate for the system control (Smith, 1957; Alevisakis and Seborg, 1973).

The term "practical stability" was first used by Palmor

(1980) to describe the ability of a time-delay system to remain stable in the presence of small perturbations. Palmor pointed out the fact that some Smith Predictor controlled systems can be destabilized by very small perturbations in the plant transfer function (matrix) coefficients and/or time-delay constants. In Section 2.1 of this paper, an example is given to show that, like Smith Predictor control, the conventional feedback control system with time-delays also have the same practical stability problem.

To ascertain the closed-loop practical stability of a multivariable control system with a Smith Predictor, Palmor and Halevi (1983) have given a *necessary* condition which has a very simple form and can be easily checked. Later Palmor and Halevi's criterion was elaborated by Yamanaka and Shimemura (1987) and developed into a *necessary and sufficient* condition for *scalar systems* test.

Based on a unified framework for time-delay system stability analysis, infinitesimal perturbations of time-delay systems are characterized and the effect on system stability is studied in this paper. A *sufficient* condition is given which can be applied to linear multivariable time-delay systems with Smith Predictors, and to conventional feedback control as well. With a modest assumption, a *necessary and sufficient* condition is derived for closed-loop practical stability of conventional feedback systems.

2. Infinitesimal perturbations and practical stability

For the systems as shown in Figs 1 and 2, the following notation is used:

$G(s)$	$(m \times r)$ transfer function matrix of the plant.
$G_m(s)$	$(m \times r)$ transfer function matrix of the plant model.
$G_0(s) \in \mathcal{R}(s)^{m \times r}$	$G(s)$ without time-delays.
$G_{m0}(s) \in \mathcal{R}(s)^{m \times r}$	$G_m(s)$ without time-delays.
$G_c(s) \in \mathcal{R}(s)^{r \times m}$	Transfer function matrix of the primary controller.

To be more specific, the element of $G(s)$, $G_0(s)$, $G_m(s)$ and $G_{m0}(s)$ are of the following form:

$$[G(s)]_{ij} = g_{ij}(s)e^{-\theta_{ij}s} \quad (1)$$

$$[G_0(s)]_{ij} = g_{ij}(s) \quad (2)$$

$$[G_m(s)]_{ij} = g_{mij}(s)e^{-\theta_{mij}s} \quad (3)$$

$$[G_{m0}(s)]_{ij} = g_{mij}(s) \quad (4)$$

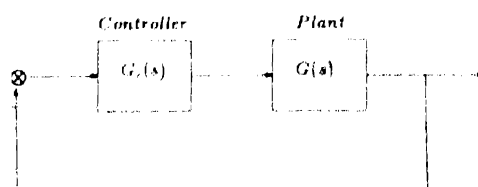


FIG. 1. Conventional feedback control.

* Received 26 January 1989; revised 2 May 1990; received in final form 15 June 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak.

† Control Systems Research, Department of Engineering, University of Leicester, Leicester LE1 7RH, U.K.

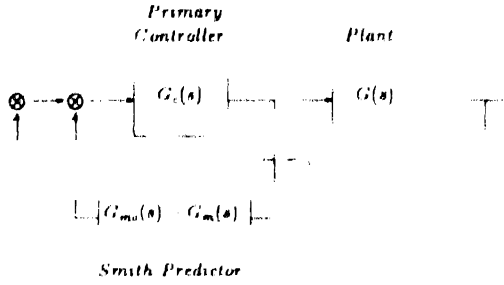


FIG. 2. Smith Predictor control.

where $g_{ij}(s), g_{mij}(s) \in \mathcal{R}(s), \theta_{ij} \geq 0, \theta_{mij} \geq 0, i = 1, 2, \dots, r, j = 1, 2, \dots, m$.

2.1. *Practical stability of conventional feedback control.* First of all, the definition of system stability given by Yamanaka and Shimemura (1987) has been adopted here and is repeated as follows.

Definition 1. Let the characteristic equation of the closed-loop system with time delays be $f(s)$ and

$$\Sigma \triangleq \{ \sigma \mid \sigma = \operatorname{Re}(z), f(z) = 0 \} \quad (5)$$

The closed-loop system is said to be stable iff Σ has a negative upper bound.

It is noted that, in general Σ is an infinite set (Bellman and Cooke, 1963; Yamanaka and Shimemura, 1987), the requirement for the negative upper bound for Σ is to exclude the case where the closed-loop poles make up an infinite chain asymptotically approaching the imaginary axis from left.

Compared with Smith Predictor control, the practical stability problem for the conventional feedback control structure has received much less attention. The following example is taken from Yamanaka and Shimemura (1987) with slight modification. It considers a fictitious system (conventional feedback control) with a seemingly negligible time-delay. It shows that, in presence of small time-delays, an infinitesimal perturbation of the system transfer function coefficients can destabilize a nominally stable (when there is no plant/model mismatch) system.

Example 1. Consider a plant with transfer function:

$$g(s) = \frac{\epsilon s^2 + \epsilon s + 1}{(\epsilon s + 1)(s + 1)(2s + 1)} e^{-\tau s}$$

where the small parameters ϵ and τ are non-negative. The control configuration is shown by Fig. 1 with the cascade compensator $g_c(s)$ as:

$$g_c(s) = \frac{1.25}{s} + 3.75 + 2.50s.$$

If the plant model is chosen to be

$$g_m(s) = \frac{1}{(s + 1)(2s + 1)}$$

then the transfer function of the nominal closed-loop system

$$\frac{1}{0.8s + 1}$$

is obviously stable. However, the open-loop transfer function of the real system is

$$q(s) = g(s)g_c(s) = \frac{1.25(\epsilon s^2 + \epsilon s + 1)}{s(\epsilon s + 1)} e^{-\tau s}.$$

It is known that the characteristic equation $1 + q(s) = 0$ has its zeros making up an infinite chain asymptotically approaching those of the comparison function

$$f_0(s) = 1 + 1.25e^{-\tau s}.$$

Consequently, the closed-loop system is practically unstable for infinitesimal τ and ϵ since $f_0(s) = 0$ has its zeros on line $\operatorname{Re}(s) = \log 1.25$ (Bellman and Cooke, 1963).

The above example shows that the stability of time-delay systems is vulnerable to even very small parameter perturbations. It has also shown that conventional feedback control has the same practical stability problem as Smith Predictor control and is a problem deserving attention.

2.2. *Practical stability of time-delay systems.* One of the advantages of applying a Smith Predictor to control a time-delay system (Fig. 2) is that, in the nominal case $G(s) = G_m(s)$, the characteristic equation of the system becomes delay-free as

$$\det [I + G_{m0}(s)G_c(s)] = 0 \quad (6)$$

Equation (6) is the same as the characteristic equation of conventional feedback control and is delay-free.

Usually, the design of $G_c(s)$ is based on $G_{m0}(s)$ so that $\det [I + G_{m0}(s)G_c(s)] \neq 0, \operatorname{Re}(s) \geq d$ which means that the system is nominally stable. In practical situations, mismatch between $G(s)$ and $G_m(s)$ always exists and can cause serious problems to closed-loop stability. Control systems which can be destabilized by a very small perturbation are of no practical value. Assume that the vector $\alpha \in \mathcal{R}^q$ denotes the coefficients of $G_0(s)$ susceptible to perturbations, and introduce $G_{m0}(s)$, where $G_{m0}(s)$ is obtained from $G_0(s)$ by setting $\alpha = \alpha_m$.

Definition 2. A control system with time-delays is termed practically stable (Palmor, 1980; Palmor and Halevi, 1983; Yamanaka and Shimemura, 1987) if:

- (1) The system is closed-loop stable for the nominal case.
- (2) There exist positive numbers δ_α and δ_θ , such that, for α and θ_j satisfying $\|\alpha - \alpha_m\| < \delta_\alpha, |\theta_j - \theta_{mj}| < \delta_\theta (\forall j)$, the closed-loop system is stable.

2.3. *A diagonalisation procedure by structural decomposition.* It is observed that, if the plant model is chosen to be delay-free such that $G_m = G_{m0}$, the Smith Predictor control (Fig. 2) is reduced to the conventional feedback control (Fig. 1). For expositional clarity, the problem formulation is based on the Smith Predictor control scheme first.

Consider the Smith Predictor control system as shown by Fig. 2. The characteristic equation of the overall closed-loop system is:

$$\begin{aligned} & \det [I_m + GG_c(I_m + (G_{m0} - G_m)G_c)^{-1}] \\ &= \det [I_m + G_{m0}G_c + (G - G_m)G_c] / \det [I_m + (G_{m0} - G_m)G_c] \\ &= \frac{\det [I_m + (G - G_m)G_c(I_m + G_{m0}G_c)^{-1}] \det [I_m + G_{m0}G_c]}{\det [I_m + (G_{m0} - G_m)G_c]} \end{aligned} \quad (7)$$

In the nominal case $G = G_m$, equation (7) is reduced to:

$$\det [I_m + G_{m0}G_c] = 0.$$

The primary controller $G_c(s)$ design ensures the closed-loop stability of the nominal system, thus, the actual overall system has a characteristic equation as:

$$\det [I_m + (G - G_m)G_c(I_m + G_{m0}G_c)^{-1}] = 0. \quad (8)$$

Assume that elementwise knowledge of perturbation on G is available and there are altogether k perturbed elements in G . Without loss of generality, order the k perturbed elements and denote them as $g_q(s)e^{-\theta_q s}$ ($q = 1, 2, \dots, k$). The corresponding elements in model G_m are $g_{mq}(s)e^{-\theta_{mq}s}$ ($q = 1, 2, \dots, k$). It is noted that a perturbation may exist only in the time constants θ_q or simultaneously in both $g_q(s)$ and θ_q ($q = 1, 2, \dots, k$).

Thus there are only k non-zero elements in $(G - G_m)$, and $(G - G_m)$ can be diagonalized by a structural decomposition as:

$$G - G_m = E \Delta G(s) F \quad (9)$$

with $E \in \mathcal{R}^{m \times k}, \Delta G(s) \in \mathcal{E}^{k \times k}$, and $F \in \mathcal{R}^{k \times m}$. Both E and F are constant matrices whose elements are 1s and 0s. $\Delta G(s)$ is a diagonal transfer function matrix and its elements are the

difference between perturbed plant elements $g_q(s)e^{-\theta_q s}$ and the corresponding model elements $g_{m_q}(s)e^{-\theta_{m_q}s}$ ($q = 1, 2, \dots, k$). Let

$$G_d = \text{diag} \{g_1(s)e^{-\theta_1 s}, \dots, g_k(s)e^{-\theta_k s}\} \quad (10)$$

$$G_{d,m} = \text{diag} \{g_{m_1}(s), \dots, g_{m_k}(s)\} \quad (11)$$

$$\tau_q = \theta_q - \theta_{m_q}, \quad q = 1, 2, \dots, k \quad (12)$$

$$S_\alpha(s) = \text{diag} \left\{ \frac{g_1(s)}{g_{m_1}(s)}, \dots, \frac{g_k(s)}{g_{m_k}(s)} \right\} \quad (13)$$

$$S_\theta(s) = \text{diag} \{e^{-\theta_{m_1}s}, e^{-\theta_{m_2}s}, \dots, e^{-\theta_{m_k}s}\} \quad (14)$$

$$S_\tau(s) = \text{diag} \{e^{-\tau_1 s}, e^{-\tau_2 s}, \dots, e^{-\tau_k s}\} \quad (15)$$

then $\Delta G(s)$ can be written as

$$\Delta G(s) = G_d - G_{d,m}S_\theta \quad (16)$$

Thus the left hand side of (8) can be rewritten as:

$$\begin{aligned} & \det [I_m + (G - G_m)G_i(I_m + G_{m0}G_i)^{-1}] \\ &= \det [I_m + E(\Delta G)FG_i(I_m + G_{m0}G_i)^{-1}] \\ &= \det [I_k + (G_d - G_{d,m}S_\theta)FG_i(I_m + G_{m0}G_i)^{-1}E] \\ &= \det [I_k + S_\theta(s)(S_\alpha S_\tau - I_k)G_{d,m}FG_i(I_m + G_{m0}G_i)^{-1}E]. \end{aligned}$$

Let

$$M(s) = G_{d,m}FG_i(s)(I_m + G_{m0}(s)G_i(s))^{-1}E \quad (17)$$

then the closed-loop system is practical stable if, and only if, the closed-loop characteristic equation

$$\det [I_k + S_\theta(s)(S_\alpha(s)S_\tau(s) - I_k)M(s)] \neq 0 \quad (18)$$

has a negative upper bound for the real parts of its zeros.

Equation (18) characterizes the practical stability of the original system with the diagonalized uncertainty $S_\theta(S_\alpha S_\tau - I_k)$, the multivariable nature of the problem is preserved by matrix M which is generally not a diagonal matrix. It is the uncertain part $S_\theta(S_\alpha S_\tau - I_k)$ which causes the problem of practical instability.

2.4. Characterization of infinitesimal perturbations. Use the same notations as in Sections 2.2 and 2.3, the perturbed coefficients of the plant are represented by the real vector $\alpha \in \mathcal{R}^p$, and its nominal value is $\alpha_m \in \mathcal{R}^p$. Perturbations in delay time constants are denoted τ_q , $q = 1, 2, \dots, k$.

In the ideal situation, $\alpha = \alpha_m$ and $\tau_q = 0$ ($q = 1, 2, \dots, k$), and hence $S_\alpha(s) = I_k$ and $S_\tau(s) = I_k$; thus, the uncertainty $S_\theta(s)(S_\alpha(s)S_\tau(s) - I_k)$ in (18) is actually a zero matrix $[0]_{k \times k}$. For any $s = \sigma + j\omega$ with $\sigma \leq d$ and $|\omega| \leq \tau$, all the elements of transfer function matrix $Q(s) \triangleq S_\theta(s)(S_\alpha(s)S_\tau(s) - I_k)M(s)$ are analytic functions of α and τ_i ($i = 1, 2, \dots, k$). Therefore, for any given δ , there exist δ_α and δ_τ such that, for all α and θ_i ($i = 1, 2, \dots, k$) satisfying $\|\alpha - \alpha_m\| \leq \delta_\alpha$, $|\tau_i| \leq \delta_\tau$ ($\forall i$), $\|Q(s)\| \leq \delta$ holds. To summarize, the following lemma applies.

Lemma 1. Assume that $M(s)$ is a stable transfer function matrix and the mismatch $S_\alpha(s)$ caused by coefficient perturbation is also stable. In the closed half plane $\text{Re}(s) \geq d$, all the zeros of equation (18) can only be at $s \rightarrow \tau$ under infinitesimal perturbations.

3. Practical stability tests for time-delay systems

By Lemma 1, the practical stability test is needed only at s which is at a sufficiently large distance from the origin. However, uncertain matrices $S_\theta(s)$, $S_\alpha(s)$ and $S_\tau(s)$ in (14), (13) and (15) present a computational difficulty. Consequently, some easily obtained criteria are highly desirable for practical stability tests. In this part of the paper, tight conditions are given which guarantee practical stability of general time-delay systems. For a special class of systems, necessary and sufficient conditions are also presented.

3.1. An assumption. Based on Lemma 1, an assumption is made on the uncertainty caused by coefficient perturbation.

Assumption 1. Assume that there exist some non-negative integer vector $N = [n_1, n_2, \dots, n_k]^T$, some non-negative

vector $\beta = [\beta_1, \beta_2, \dots, \beta_k]^T$, and $r > 0$, such that

$$\left| \frac{g_q(s)}{g_{m_q}(s)} - 1 \right| \leq b_q |s|^{\alpha_q}, \quad \text{for } \text{Re}(s) \leq d \quad (19)$$

$$\text{and } |\omega| \leq r, \quad q = 1, 2, \dots, k$$

and

$$S(s) = \text{diag} \{b_1 s^{\alpha_1}, b_2 s^{\alpha_2}, \dots, b_k s^{\alpha_k}\}. \quad (20)$$

In the light of the Lemma 1, the above assumption is a fairly reasonable one and occurs in most practical situations which is also shown by the given example. In the finite region of the s plane with $\text{Re}(s) \leq d$ and $|\omega| \leq r$, mismatch between $g_q(s)$ and $g_{m_q}(s)$ by small coefficient perturbation is too small to have any effect on system stability and, at a distance far away from the origin, $|g_q(s)/g_{m_q}(s) - 1|$ approaches $b_q |s|^{\alpha_q}$. By restricting the perturbation to arbitrarily small, the distance r over which $|g_q(s)/g_{m_q}(s) - 1|$ is negligible becomes arbitrarily large.

Based upon the above assumption, some criteria are derived which reduce the practical stability tests into a structured singular value computation for known matrices.

3.2. Practical stability tests for general systems. Since Palmor and Halevi's (1983) original conclusion (Theorem 2, p. 258) does not consider the uncertainty factor $S(s)$ caused by small coefficient perturbations, there can be a considerable gap between their test and the exact condition in the presence of coefficient perturbation. The following theorem seeks to reduce the gap.

Theorem 1 (General time-delay systems) With the assumption (19) and (20), the closed-loop system with time-delays is practical stable if

$$\mu\{(S(j\omega)^* + 2I_k)M(j\omega)\} < 1, \quad \text{as } \omega \rightarrow \infty \quad (21)$$

The structured singular value corresponds to the diagonal uncertainty structure

Proof. See Appendix 1.

It is noted that Theorem 1 can be equally applied to conventional feedback systems as well as Smith Predictor controlled systems. The computation of the structured singular value (Doyle, 1982) involved in (21) is needed only once with ω being sufficient large.

Theorem 1 is derived by inspecting the eigenvalues of matrix $Q(j\omega) = S_\theta(j\omega)(S_\alpha(j\omega)S_\tau(j\omega) - I_k)M(j\omega)$ in (18) as $\omega \rightarrow \infty$. $S_\theta(j\omega)$ and $S_\tau(j\omega)$ are diagonal matrices and their elements are pure time delays of different constants. At sufficiently high frequencies, $S_\theta(j\omega)$ and $S_\tau(j\omega)$ can almost always provide phase shifts to make

$$\begin{aligned} & \min \{ \text{Re} \lambda_i [S_\theta(j\omega)(S_\alpha(j\omega)S_\tau(j\omega) - I_k)M(j\omega)] \} \\ &= -\rho \{ [S(j\omega)^* + I_k]M(j\omega) \} \end{aligned}$$

and the diagonal structure of $S_\theta(j\omega)$ suggests the replacement of spectral radius ρ by the structured singular value μ , thus the sufficient condition (21) is expected to be fairly close to the exact condition.

However, in the case of Smith Predictor control and where the perturbation exists in only one plant element with $k = 1$, $S_\theta(j\omega)$, $S_\alpha(j\omega)$, $S_\tau(j\omega)$, and $M(j\omega)$ are scalars. With $S_\theta(j\omega)$ providing arbitrary phase shift at high frequencies, the scalar counterpart of inequality (21) becomes the exact condition for system practical stability.

Corollary 1 (Smith Predictor control). In the case of Smith Predictor control and where there is only one plant element which is perturbed, for example as in a scalar system, the closed-loop system is practical stable if, and only if,

$$(S(j\omega)^* + 2) |M(j\omega)| < 1, \quad \text{as } \omega \rightarrow \infty. \quad (22)$$

Proof. See Appendix 2.

It is noted that Corollary 1 can only be used in Smith Predictor control systems with only one perturbed element ($\theta \neq \theta_m > 0$). If Corollary 1 is applied to a scalar Smith Predictor control system, there is no need for structural

decomposition and the transfer matrix $M(s)$ reduces to the complementary sensitivity transfer function:

$$M(s) = \frac{g_{m0}(s)g_c(s)}{1 + g_{m0}(s)g_c(s)}$$

Corollary 1 turns out to be the necessary and sufficient condition given by Yamanaka and Shimemura (1987, Theorem 1, p. 789).

3.3. *Practical stability test for a class of conventional feedback control systems.* A conventional feedback control system has the following features:

- (a) $S_0(s) \approx I_k$
- (b) $G_m(s) \approx G_{m0}(s)$

Consider a class of systems satisfying:

$$\det[I_k - M(s)] \neq 0, \quad \text{Re}(s) \geq d. \tag{23}$$

The exact test for practical stability of this class of systems has a very neat form and is summarized in the following theorem:

Theorem 2 (Conventional feedback control). For the class of conventional feedback systems satisfying condition (23) together with Assumption 1, the closed-loop system is practical stable if, and only if,

$$\mu[(S(j\omega)^{-1} + I)M(j\omega)(I_k - M(j\omega))^{-1}] < 1, \quad \text{as } \omega \rightarrow \infty \tag{24}$$

where the structured singular value μ corresponds to the diagonal structure.

Proof. See Appendix 3.

It should be emphasized that Theorem 2 can only be applied to conventional feedback control systems. It should also be pointed out that a very large class of systems may satisfy condition (23). For example, if perturbations occur at the plant outputs (or inputs) and can be written in a multiplicative form, then there will be no need for structural decomposition and $M(s)(I_k - M(s))^{-1} = G_{m0}(s)G_c(s)$ (or $G_c(s)G_{m0}(s)$).

If Theorem 2 is applied to a scalar conventional feedback control system, the closed-loop system is practical stable if, and only if,

$$(S(j\omega)^{-1} + 1)|g_c(j\omega)g_m(j\omega)| < 1, \quad \text{as } \omega \rightarrow \infty. \tag{25}$$

If Corollary 1 is applied to a scalar Smith Predictor controlled system with the same perturbation, the closed-loop system is practical stable if, and only if,

$$(S(j\omega)^{-1} + 2) \frac{g_c(j\omega)(g_m(j\omega))}{1 + g_c(j\omega)g_m(j\omega)} < 1 \quad \text{as } \omega \rightarrow \infty. \tag{26}$$

It can be easily proved that inequality (26) implies inequality (25), but not *vice-versa*. So on one hand, it is emphasized that conventional feedback systems, just like Smith Predictor controlled systems, have the same practical stability problem in the presence of infinitesimal time-delays. However, it should also be pointed out that conventional feedback control can be slightly less sensitive to infinitesimal perturbations than the Smith Predictor control with the same perturbations.

4 Examples

Example 2. Consider a Smith Predictor controlled system with the plant, plant model and the primary controller as:

$$G(s) = \begin{bmatrix} e^{-0.1s} & e^{-0.25s} \\ s+1 & s+1 \end{bmatrix}$$

$$G_m(s) = \begin{bmatrix} s+1 & e^{-0.25s} \\ s+1 & s+1 \end{bmatrix}$$

$$G_c(s) = \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$$

where τ is a small perturbation parameter. The plant model without delay terms is:

$$G_{m0}(s) = \begin{bmatrix} 1 & \frac{1}{s+1} \\ \frac{1}{s+1} & \frac{1}{s+1} \end{bmatrix}$$

For the plant, the only element that is perturbed is the (1, 1) element. To isolate (diagonalize) the perturbed element, let

$$E = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad F = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

then

$$G(s) - G_m(s) = E(e^{-0.1s} - 1)F$$

$$G_{d0}(s) = 1$$

$$S_0(s) = 1$$

$$S_0(s) = e^{-0.1s}$$

$$S(s) = 0$$

$$M(s) = G_{d0}(s)FG_c(s)(I_m + G_{m0}(s)G_c(s))^{-1}E \\ = \frac{k_1(s+1)(s+k_2+1)}{(k_1+1)s^2 + (k_1+1)(k_2+2)s + (k_1+k_2+1)}$$

By Corollary 1, the closed-loop system is practically stable if, and only if,

$$(2 + S(s)^{-1})|M(s)| < 1 \quad \text{as } |s| \rightarrow +\infty$$

that is

$$2k_1/(k_1+1) < 1$$

then

$$k_1 < 1.$$

So the closed-loop system is practically stable if, and only if, $k_1 < 1$.

Example 3. Consider a 2-input, 2-output time-delay system with conventional feedback control. The plant and plant model are:

$$G(s) = \begin{bmatrix} (\epsilon s^2 + \epsilon s + 1)e^{-\tau_1 s} & \frac{1}{s+1} & \frac{0.5}{2s+1} \\ \frac{\epsilon s}{\epsilon s + 1} & 0.5 & 1 \\ 0 & \frac{1}{2s+1} & \frac{1}{s+1} \end{bmatrix}$$

and

$$G_m(s) = \begin{bmatrix} 1 & 0.5 \\ s+1 & 2s+1 \\ 0.5 & 1 \\ 2s+1 & s+1 \end{bmatrix}$$

where ϵ , τ_1 , and τ_2 are small perturbations. The controller is:

$$G_c(s) = \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$$

It can be seen that the plant perturbation occur in output channels and is already in diagonal multiplicative form, hence:

$$E = I_2, \quad F = I_2,$$

$$G_{d0} = G_{m0} = G_m = \begin{bmatrix} 1 & 0.5 \\ s+1 & 2s+1 \\ 0.5 & 1 \\ 2s+1 & s+1 \end{bmatrix}$$

$$S_0 = \begin{bmatrix} \frac{\epsilon s^2 + \epsilon s + 1}{\epsilon s + 1} & 0 \\ 0 & 1 \end{bmatrix}, \quad S_0 = I, \quad S_\tau = \begin{bmatrix} e^{-\tau_1 s} & 0 \\ 0 & e^{-\tau_2 s} \end{bmatrix}$$

$$\begin{aligned} M(s) &= G_{dmd}(s)FG_c(I_2 + G_m(s)G_c)^{-1}E \\ &= G_m(s)G_c(I_2 + G_m(s)G_c)^{-1} \\ I - M(s) &= (I_2 + G_m(s)G_c)^{-1} \end{aligned}$$

and

$$M(s)(I - M(s))^{-1} = G_m(s)G_c$$

It is obvious that for $k_1 > 0$ and $k_2 > 0$, the nominal system is stable and $\det[I - M(s)] \neq 0$, $\text{Re}(s) \geq d$, namely, Condition (23) is satisfied. As $|s| \rightarrow +\infty$,

$$\frac{\epsilon s^2 + \epsilon s + 1}{\epsilon s + 1} \rightarrow 1$$

thus

$$S(s) = \text{diag}\{s, 0\}$$

and

$$\begin{aligned} M(s)(I_k - M(s))^{-1} &= G_m(s)G_c \\ &= \begin{bmatrix} 1 & 0.5 \\ s+1 & 2s+1 \end{bmatrix} \\ &= \begin{bmatrix} 0.5 & 1 \\ 2s+1 & s+1 \end{bmatrix} \end{aligned}$$

As $\omega \rightarrow$

$$\begin{aligned} (S(j\omega)^* + I)G_m(j\omega)G_c &= \begin{bmatrix} k_1 & 0.25k_2 \\ 0 & 0 \end{bmatrix} \\ \mu \left(\begin{bmatrix} k_1 & 0.25k_2 \\ 0 & 0 \end{bmatrix} \right) &= k_1 \end{aligned}$$

By Theorem 2, the closed-loop system is practically stable if, and only if, $k_1 < 1$.

5. Conclusions

Based on the framework of Smith Predictor control in which the conventional feedback is a special case, the practical stability problem of time-delay systems has been investigated. It is revealed that the practical stability problem can exist in conventional feedback systems as well as in Smith Predictor control systems. Criteria have been established which guarantee practical stability of linear multivariable time-delay systems under infinitesimal perturbations in plant coefficients and delay time constants. The proposed practical stability tests can be easily performed and provide guidelines for the control design of time-delay systems.

Acknowledgement—The research reported in this paper was supported by a British Council postgraduate scholarship. The author wishes to thank Dr M. A. Johnson (University of Strathclyde) for his constant encouragement during the research and Prof. I. Postlethwaite (University of Leicester) for providing the opportunity to finish the paper.

References

- Alevisakis, G. and D. E. Seborg (1973). An extension of the Smith Predictor method to multivariable linear systems containing time delays. *Int. J. Control*, **3**, 543–551.
- Bellman, R. and K. L. Cooke (1963). *Differential-Difference Equations*. Academic Press, New York.
- Doyle, J. C. (1982). Analysis of feedback systems with structured uncertainties. *Proc. IEE*, Pt D, **129**, 242–250.
- Palmer, Z. (1980). Stability properties of Smith dead-time compensator controllers. *Int. J. Control*, **32**, 937–949.
- Palmer, J. Z. and Y. Halevi (1983). On the design and properties of multivariable dead time compensators. *Automatica*, **19**, 255–264.
- Smith, O. J. M. (1957). A controller to overcome dead time. *ISA J.*, **6**, p. 28.
- Takahashi, S., K. Yamanaka and M. Yamada (1987). Detection of dominant poles of systems with time delay by using Pade approximation. *Int. J. Control*, **45**, 251–254.

Yamanaka, K. and E. Shimemura (1987). Effects of mismatched Smith controller on stability in systems with time-delay. *Automatica*, **23**, 787–791.

Appendix 1 (Proof of Theorem 1)

As frequency ω approaches infinity, some properties of the pure time-delay diagonal matrices $S_s(j\omega)$ and $S_\theta(j\omega)$ will be used in the theorem proof and are listed as follows:

- (a) At a given sufficiently high frequency band, each diagonal element of $S_s(j\omega)$ can provide arbitrarily independent pure phase shift.
- (b) Even if the time delay constants of $S_\theta(s)$ are distinct, at any frequency point ω , the phase shifts provided by the diagonal elements of $S_\theta(j\omega)$ are dependent. But at some sufficiently high frequency points, the phase shifts by $S_\theta(j\omega)$ can be close to some required phase shifts.

The proof of the theorem can be carried out in two steps.

- (1) Let $s_\omega = \sigma + j\omega$ with $\sigma \geq d$ and $|s_\omega| \rightarrow +\infty$, it can be proved that, if $\sigma \geq 0$ and $\mu[S(j\omega)^* + 2I_k]M(j\omega) < 1$ as $\omega \rightarrow \infty$, equation (18) has no zeros for $s = s_\omega$.

First, the diagonal matrix

$$X(s) = S_\theta(s)(S_s(s)S_\theta(s) - I_k)(S(s)^* + 2I_k)^{-1}$$

with $s = s_\omega$ and $\sigma \geq 0$ is investigated where $S_s(s)$, $S_\theta(s)$, $S(s)$, and $S(s)$ are given in (13), (14), (15), and (20) respectively. As $|s_\omega| \rightarrow +\infty$, the absolute value of the (i, i) th element of $X(s_\omega)$ is:

$$\begin{aligned} & \mu_{ii} \left(\frac{e^{-\sigma_\theta |s_\omega|} g_i(s_\omega)}{g_{mi}(s_\omega)} - 1 \right) \\ & \quad \frac{b_i |s_\omega|^{n_i} + 2}{e^{-\sigma |s_\omega|} \left(\frac{g_i(s_\omega)}{g_{mi}(s_\omega)} - 1 \right) + |e^{-\sigma |s_\omega|} - 1|} \\ & \quad \frac{b_i |s_\omega|^{n_i} + 2}{\left| \frac{g_i(s_\omega)}{g_{mi}(s_\omega)} - 1 \right| + 2} \\ & \quad \frac{b_i |s_\omega|^{n_i} + 2}{b_i |s_\omega|^{n_i} + 2} \quad (\text{by (19)}) \\ & \quad \leq 1 \end{aligned}$$

Consequently, the largest singular value of $X(s_\omega)$ satisfies $\sigma_{\max}[X(s_\omega)] < 1$ and

$$\begin{aligned} & \rho[S_\theta(s_\omega)(S_s(s_\omega)S_\theta(s_\omega) - I_k)M(s_\omega)] \\ & \quad = \rho[X(s_\omega)(S(s_\omega)^* + 2I_k)M(s_\omega)] \\ & \quad \leq \mu[(S(s_\omega)^* + 2I_k)M(s_\omega)] \end{aligned}$$

therefore, equation (18) has no zeros for $\text{Re}(s_\omega) \geq 0$ if $\mu[(S(s_\omega)^* + 2I_k)M(s_\omega)] < 1$ with $|s_\omega| \rightarrow +\infty$, which is the same as $\mu[S(j\omega)^* + 2I_k]M(j\omega) < 1$ with $\omega \rightarrow \infty$.

- (2) In what follows, it is going to prove that, if $\mu[S(j\omega)^* + 2I_k]M(j\omega) < 1$ with $\omega \rightarrow \infty$, there will exist a $\sigma_0 < 0$ such that, equation (18) has no zeros for $s = s_\omega$ and $\sigma_0 \leq \sigma \leq 0$.

Let $\mu[(S(s_\omega)^* + 2I_k)M(s_\omega)] < \mu_0 < 1$ and investigate the following function:

$$h_i(\sigma) = \frac{g_i(s_\omega)}{g_{mi}(s_\omega)} - 1 + |e^{-\sigma |s_\omega|} - 1| \frac{b_i |s_\omega|^{n_i} + 2}{b_i |s_\omega|^{n_i} + 2}$$

Obviously, for $\sigma \leq 0$, $h_i(\sigma) \leq 1$ is monotonically decreasing with σ . Let σ_{i0} be the unique solution of

$$h_i(\sigma_{i0}) = 1$$

then $h_i(\sigma) < 1/\mu_0, \forall \sigma_0 < \sigma \leq 0$. Define:

$$\sigma_0 = \max \{ \sigma_{10}, \sigma_{20}, \dots, \sigma_{k0} \}.$$

It follows that, for $\sigma_0 \leq \sigma \leq 0$, the absolute value of the (i, i) th element of diagonal matrix $X(s_\omega)$ is:

$$\begin{aligned} & \left| \frac{e^{-\sigma_0} \left(e^{-\sigma_0} \frac{g_i(s_\omega)}{g_{m_i}(s_\omega)} - 1 \right)}{b_i |s_\omega|^{n_i} + 2} \right| \\ & \leq \frac{n_i \sigma \left(e^{-\sigma_0} \frac{g_i(s_\omega)}{g_{m_i}(s_\omega)} + e^{-\sigma_0} + 1 \right)}{b_i |s_\omega|^{n_i} + 2} \\ & = h_i(\sigma) \\ & \leq h_i(\sigma_0) \\ & \leq \frac{1}{\mu_0} \end{aligned}$$

which means that the diagonal matrix $X(s_\omega)$ has its largest singular value $\sigma_{\max}[X(s_\omega)] \leq 1/\mu_0$. On the other hand,

$$\begin{aligned} & \rho[S_0(s_\omega)(S_\alpha(s_\omega)S_\tau(s_\omega) - I_k)M(s_\omega)] \\ & = \rho[X(s_\omega)(S(s_\omega)^* + 2I_k)M(s_\omega)] \\ & \leq \sigma_{\max}[X(s_\omega)]\mu[(S(s_\omega)^* + 2I_k)M(s_\omega)] \\ & \leq \frac{\mu[(S(s_\omega)^* + 2I_k)M(s_\omega)]}{\mu_0} \\ & < 1. \end{aligned}$$

Therefore, equation (18) has no zeros for $\text{Re}(s_\omega) < \sigma_0$ if $\mu[(S(s_\omega)^* + 2I_k)M(s_\omega)] < 1$.

Combine (1) and (2) with Lemma 1, it is obtained that equation (18) has no zeros for $\text{Re}(s) < \sigma_0$ if $\mu[(S(j\omega)^* + 2I_k)M(j\omega)] < 1$ as $\omega \rightarrow \infty$. In other words, σ_0 can be regarded as the negative upper bound for Σ defined in Definition 1, and the closed-loop system is practically stable.

Appendix 2 (Proof of Corollary 1)

In this case, the sufficiency of the condition (22) can be directly obtained from Theorem 1. The proof of necessity is as follows.

Assume that $|(S(j\omega)^* + 2)M(j\omega)| \geq 1$, from the proof of Theorem 1, a non-negative scalar $\sigma_0 \geq 0$ can be found such that

$$h(\sigma_0) = \frac{1}{|(S(j\omega)^* + 2)M(j\omega)|} \leq 1.$$

What is more is that $e^{-j\sigma_0\omega}$ can give $X(\sigma_0 + j\omega)$ an arbitrary phase. Thus, $X(\sigma_0 + j\omega)$ can provide the exact magnitude scaling as well as phase shift to make

$$\begin{aligned} & S_0(\sigma_0 + j\omega)(S_\alpha(\sigma_0 + j\omega)S_\tau(\sigma_0 + j\omega) - I)M(\sigma_0 + j\omega) \\ & = X(\sigma_0 + j\omega)(S(j\omega)^* + 2)M(j\omega) = -1 \end{aligned}$$

at certain discrete frequency points with $\omega \rightarrow \infty$, which means that the equation (18) has a infinite chain of zeros on line $\text{Re}(s) = \sigma_0 \geq 0$, therefore the closed-loop system is not practically stable. Consequently, Condition (22) is also necessary.

Appendix 3 (Proof of Theorem 2)

In the case $S_0(s) = I_k$, with $s_\omega = \sigma + j\omega$ and $|s_\omega| \rightarrow +\infty$,

$$\det[I_k + (S_\alpha(s_\omega)S_\tau(s_\omega) - I_k)M(s_\omega)] \neq 0, \quad \text{Re}(s_\omega) \geq d$$

is equivalent to

$$\begin{aligned} & \det[I_k + S_\tau(s_\omega)S_\alpha(s_\omega)M(s_\omega)(I_k - M(s_\omega))^{-1}] \neq 0, \\ & \text{Re}(s_\omega) \geq d \end{aligned}$$

under the assumption (23). This time, define:

$$X(s) \triangleq S_\alpha(s)(S(s)^* + I)^{-1}$$

and keep in mind that $S_\tau(j\omega)$ can provide arbitrary phase shifts, the proof of Theorem 2 can be carried out in the same way as the combined Theorem 1 proof and Corollary 1 proof.

Brief Paper

Implied Polynomial Matrix Equations in Multivariable Stochastic Optimal Control†

K. J. HUNT‡ and M. ŠEBEK§

Key Words—Algebraic system theory; stochastic optimal control theory; multivariable control systems; polynomial equation approach.

Abstract—This paper reports recent work in the theoretical development of the polynomial equation approach to the optimization of multivariable control systems. The algebraic properties of the polynomial matrix equations which define the optimal controller are investigated, and new results concerned with the numerical solvability of the equations are derived.

Notation—All systems considered in this paper are described by means of real polynomial matrices in the delay operator d . The reader is referred to Kučera (1979) for details. For simplicity the arguments of polynomial matrices are often omitted, such that $X(d)$ is denoted by X . The adjoint of $X(d)$ is written as $X^*(d) = X'(d^{-1})$. For any polynomial matrix $X(d)$ define $\langle X \rangle$ as the matrix of terms independent of d . Stable square polynomial matrices are those with zeros of their determinant strictly outside the unit circle of the d -plane.

1. Introduction

THE DESIGN of optimal controllers for multivariable plants subject to noise disturbances has been intensively studied in recent years. If only the plant output can be measured it is well known that the optimal controller consists of linear output feedback and can be designed using either time-domain (Kwakernaak and Sivan, 1972) or frequency-domain (Youla *et al.*, 1976) methods.

Alternatively, the optimal multivariable controller may be designed using the *polynomial equation approach* which was developed by Kučera (1979). Kučera's output regulation solution was extended to the tracking case by Šebek (1983).

If, in addition to the plant output and reference signal, some disturbance can be measured, then a three-input controller, utilizing feedback, reference control and *disturbance measurement feedforward* may be used to improve the controller performance (i.e. to decrease the optimal cost).

For the case of single-input single-output plants, the solution of such a modified optimal control problem was first given by Grimble (1986) for *stable* disturbances. The solution for possibly unstable disturbances was subsequently given by Šebek *et al.* (1988). Independently, Sternad and Söderström (1988) obtained the solution for the stable disturbance case using an alternative proof technique. A completely general solution to the scalar feedback/feedforward stochastic

tracking problem incorporating dynamic cost-function weights has been obtained by Hunt (1988). The direct feedback/feedforward regulator solution of Šebek *et al.* (1988) was recently extended to *multivariable* plants by Hunt and Šebek (1989a). An indirect solution to the multivariable feedback/feedforward problem has previously been obtained by Grimble (1988). Grimble's solution was obtained by reformulating the plant model and optimal control problem as an equivalent, augmented, output regulation problem. A full treatment for both scalar and multivariable systems is given in Hunt (1989).

The theoretical developments put forth by Hunt and Šebek (1989a) were recently used as the basis of a Multivariable LQG Self-tuning Controller Algorithm with Disturbance Measurement Feedforward (Hunt and Šebek, 1989b).

LQG self-tuning control for single-input single-output systems has previously been considered by Grimble (1984) and Hunt *et al.* (1986, 1987a), and is most fully described by Hunt (1989). Hunt (1989) also reports a detailed application study of LQG self-tuning control relating to the steam temperature control loop on an Advanced Gas-cooled Reactor power installation.

The polynomial equation approach has been shown to provide a sound basis for the application of high performance controllers and we continue here with important aspects of its development.

The optimal multivariable controller consists of three parts: a feedback part, a reference tracking part, and a feedforward part. In general, calculation of the optimal controller requires the solution of three pairs of *coupled two-sided* polynomial matrix equations, one coupled pair of equations being connected with each part of the controller. Elimination of the common (or coupling) term between the coupled equations results in three *single one-sided* polynomial matrix equations, the so-called *implied* polynomial matrix equations. The minimum degree solution of the original three pairs of coupled equations results in the optimal controller, which shifts the poles and zeros of the closed-loop system to their desired optimal positions. On the other hand, a controller calculated using the minimum degree solution of the implied equations ensures only the optimal positions of the closed-loop poles. The conditions under which the implied *feedback equation does* yield the optimal *feedback* controller have been derived by Hunt *et al.* (1987b). The related conditions for the tracking controller equations and the feedforward controller equations are established and proven in this paper.

This question is of crucial importance in the multivariable LQG self-tuning controller algorithm (Hunt and Šebek, 1989b) since the computational burden presented by solution of the single-sided implied equations is much less than that of the original two-sided coupled pairs of equations.

The role of the coupled polynomial matrix equations arising in the optimal control problem solution has also been considered by Roberts and Newmann (1988). Roberts and Newmann obtain the conditions, for *stable* plants, under which the first of the coupled equations is itself sufficient in the determination of the optimal controller.

† Received 28 November 1989; revised 11 June 1990; received in final form 11 July 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor P. M. G. Ferreira under the direction of Editor H. Kwakernaak.

‡ Department of Mechanical Engineering, University of Glasgow, Glasgow G12 8QQ, Scotland, U.K. Author to whom all correspondence should be addressed.

§ Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, Pod vodarenskou věží 4, 182 08 Prague 8, Czechoslovakia.

2. System model and cost-function

The multi-input multi-output plant under consideration is governed by the equation:

$$Ay = Bu + C_1\delta_1 + C_2\delta_2 \quad (1)$$

where y is the vector output sequence, u is the vector control input sequence, and δ_1 and δ_2 are two vector noise sequences. A , B , C_1 and C_2 are polynomial matrices in d . The plant is assumed strictly causal, so that $\langle A \rangle$ is invertible while $\langle B \rangle = 0$. The noise component δ_2 passes through a filter to produce a measured disturbance signal δ_1 , i.e.:

$$A_1\delta_1 = C_1\delta_2 \quad (2)$$

where A_1 and C_1 are polynomial matrices in d , with C_1 square. The filter $A_1^{-1}C_1$ typically represents measurement dynamics.

Further, consider a reference vector sequence r modelled as:

$$A_r r = C_r \delta_r \quad (3)$$

where A_r and C_r are left coprime polynomial matrices in d , with $\langle A_r \rangle$ invertible. δ_r is a stochastic generator vector. The available version of the reference sequence is corrupted by an additive observation noise δ_m .

The general linear controller which operates on the plant output (corrupted by a measurement noise δ_1), on the reference signal (corrupted by δ_m), and on the measured disturbance signal δ_1 is described by:

$$Pu = -Q(y + \delta_1) + R(r + \delta_m) + S\delta_1 \quad (4)$$

where P , Q , R and S are the polynomial matrices to be found, and $\langle P \rangle$ is invertible. The overall system structure is shown in Fig. 1. Note that in practice the controller must be realized as a single dynamical system having three vector inputs and one vector output (i.e. the control signal u).

All the vector random sources δ_1 , δ_2 , δ_r , δ_m and δ_m are mutually independent stationary white noises with intensities Φ_1 , Φ_2 , Φ_r , Φ_m and Φ_m , respectively. To avoid the trivial case of $\Phi_2 = 0$ (i.e. no measurable disturbance) we assume here, without loss of generality, that $\Phi_2 = I$. Φ_1 , Φ_r , Φ_m and Φ_m are real non-negative definite matrices.

The desired optimal controller evolves from minimization of the cost-function:

$$J = \text{tr}(\Omega\phi_u) + \text{tr}(\Sigma\phi_{r-y}) \quad (5)$$

where ϕ_u and ϕ_{r-y} are correlation functions of u and the tracking error $r - y$ in steady-state, respectively. Ω and Σ are real non-negative definite weighting matrices. Thus, the design problem is to minimize the cost (5) subject to the constraint that the closed-loop system defined by equations (1)–(4) be asymptotically stable.

3. Optimal controller

The first stage in the design procedure is to find a pair of right-coprime polynomial matrices A_1 and B_1 and left-coprime polynomial matrices A_0 and B_0 such that:

$$A^{-1}B = B_1A_1^{-1} = A_0^{-1}B_0 \quad (6)$$

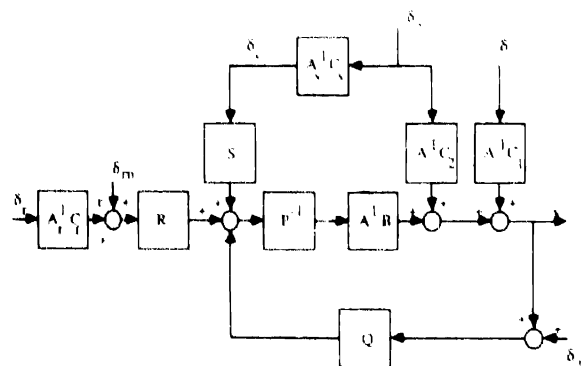


FIG. 1. Closed-loop system.

We define stable polynomial matrices D_c , D_f and G (the spectral factors) which satisfy:

$$A_1^* \Omega A_1 + B_1^* \Sigma B_1 = D_c^* D_c \quad (7)$$

$$A \Phi_1 A^* + C_1 \Phi_1 C_1^* = D_f^* D_f \quad (8)$$

$$A \Phi_m A^* + C_r \Phi_r C_r^* = G G^* \quad (9)$$

For brevity we assume here that the given data make the problem regular, and that stable spectral factors do exist. This is normally the case in practice and ensures a solution to the optimal control problem. Even when the spectral factors are not strictly stable a solution to the problem may still exist. See Kučera (1979) for a formal treatment of this case. We then calculate any left-coprime matrix fraction

$$A_4^{-1}A_1 = A_n A_r^{-1} \quad (10)$$

Further, the following right-coprime matrix fractions are defined by:

$$D_f^{-1}A = A_d D_{fa}^{-1} \quad (11)$$

$$D_f^{-1}B = B_d D_{fb}^{-1} \quad (12)$$

$$A^{-1}C_2 = C_d A_r^{-1} \quad (13)$$

$$DE^{-1} = (A_1 G)^{-1} A_4 B_0 \quad (14)$$

$$A_1 G_2^{-1} = G^{-1} A_r \quad (15)$$

Finally, we define the right-coprime polynomial matrices B_c , C_b by:

$$BC_b = C_2 B_1 \quad (16)$$

The optimal controller equations may now be stated as follows:

Theorem 1. The optimal control problem is solvable if and only if:

- The greatest common left divisor of A and B is a stable polynomial matrix.
- C_1 is a stable polynomial matrix.
- A_4 is a stable polynomial matrix.

The optimal controller polynomial matrices P , Q , R and S are obtained from the following left-coprime matrix fraction:

$$\begin{aligned} \hat{D}_f^{-1}[P, Q, R, S] \\ = [X D_{fb}^{-1}, Y D_{fa}^{-1}, M G_2^{-1}, (Y D_{fa}^{-1} C_d - Z) A_r^{-1} C_1^{-1} A_r] \end{aligned} \quad (17)$$

Here, X and Y (along with V) is the solution of the equations:

$$D_c^* X + V^* B_d = A_1^* \Omega D_{fb} \quad (18a)$$

$$D_c^* Y - V^* A_d = B_1^* \Sigma D_{fa} \quad (18b)$$

such that $\langle V \rangle = 0$.

The polynomial matrix M (along with L and N) is the solution of the equations:

$$D_c^* L + N^* D = A_1^* \Omega E \quad (19a)$$

$$D_c^* M - N^* A_r = B_1^* \Sigma G_2 \quad (19b)$$

such that $\langle N \rangle = 0$.

The polynomial matrix Z (along with U and W) is the solution of the equations:

$$D_c^* U + W^* B_c = A_1^* \Omega C_b \quad (20a)$$

$$D_c^* Z - W^* A_r = B_1^* \Sigma C_d \quad (20b)$$

such that $\langle W \rangle = 0$.

Proof. The feedback part of the optimal controller (i.e. relating to P and Q) was first derived by Kučera (1979). The tracking part (R) was subsequently obtained by Šebek (1983). Finally, the feedforward part (S) of the theorem was derived by Hunt and Šebek (1989a).

4. Implied polynomial matrix equations

Lemma. The polynomial matrices X and Y in equation (18) also satisfy the implied feedback polynomial matrix equation:

$$YB_1 + XA_1 = D_1 D_2 \quad (21)$$

where

$$\begin{bmatrix} D_{1a}^{-1} B_1 \\ D_{1b}^{-1} A_1 \end{bmatrix} = \begin{bmatrix} B_1 \\ A_1 \end{bmatrix} D_1^{-1} \quad (22)$$

Proof. Eliminating V^* from equation (18) by multiplication, adding and some algebraic manipulation, results in (21).

Lemma. When

(C.1) A_1 is a right divisor of A_0
then the polynomial matrices L and M in equation (19) also satisfy the implied reference polynomial matrix equation:

$$LA_1 + MB_1 = D_1 D_2 \quad (23)$$

where:

$$\begin{bmatrix} E^{-1} A_1 \\ G_2^{-1} B_1 \end{bmatrix} = \begin{bmatrix} A_1 \\ B_1 \end{bmatrix} D_1^{-1} \quad (24)$$

Proof. Eliminating N^* from (19) by multiplication, adding and some algebraic manipulation, results in (23).

Lemma. The polynomial matrices U and Z in (20) also satisfy the implied feedforward polynomial matrix equation:

$$UA_2 + ZB_2 = D_1 D_2 \quad (25)$$

where:

$$\begin{bmatrix} A_1^{-1} A_1 \\ A_1^{-1} C_b \end{bmatrix} = \begin{bmatrix} B_2 \\ D_2 \end{bmatrix} A_1^{-1} \quad (26)$$

Proof. Eliminating W^* from (20) by multiplication, adding and some algebraic manipulation results in (25).

Theorem 2 (main result). Solution of the implied polynomial matrix equations (21), (23) and (25) under the conditions

$$\begin{aligned} YA_d^{-1} &\text{ strictly proper} \\ MA_1^{-1} &\text{ strictly proper} \\ ZA_1^{-1} &\text{ strictly proper} \end{aligned}$$

will uniquely determine the optimal controller matrices if and only if

- (C.2) $A_1^{-1} C_a$, $A_1^{-1} C_1$, and $A_1^{-1} C_2$ are proper rational matrices
- (C.3) The polynomial matrices A and B are left coprime.

Proof. A proof of the result for the equations relating to the feedback part of the controller has been given by Hunt *et al.* (1987b). Proof of the results for the tracking and feedforward equations follows:

(a) **Tracking equations.** The optimal solution is characterized by:

$$\langle N \rangle = 0 \Leftrightarrow D_1^{-1} N^* \text{ strictly proper.}$$

After some straightforward manipulation equation (19b) can be written as:

$$MA_1^{-1} - D_1^{-1} N^* = D_1^{-1} B_1^* \Sigma A_1^{-1} G \quad (27)$$

By condition (C.2) and equation (9) $A_1^{-1} G$ is proper. Since $\langle B_1 \rangle = 0$, $D_1^{-1} B_1^*$ is strictly proper. Thus, together with $\langle N \rangle = 0$, equation (27) implies that MA_1^{-1} is strictly proper. Thus, when (C.2) holds the optimal solution is also characterized by

$$MA_1^{-1} \text{ strictly proper.}$$

Now, when A and B are left coprime (condition C.3) then $A_1^{-1} D = B_1 A_1^{-1}$ is coprime. Thus, using Lemma A1 (see Appendix) there exists a unique solution L , M to equation (23) with MA_1^{-1} strictly proper. This unique solution, as shown above, is the optimal solution.

(b) **Feedforward equations.** The optimal solution is characterized by:

$$\langle W \rangle = 0 \Leftrightarrow D_1^{-1} W^* \text{ strictly proper.}$$

After some manipulation equation (20b) can be written as:

$$ZA_1^{-1} - D_1^{-1} W^* = D_1^{-1} B_1^* \Sigma A_1^{-1} C_2 \quad (28)$$

By condition (C.2) $A_1^{-1} C_2$ is proper. Since $\langle B_1 \rangle = 0$, $D_1^{-1} B_1^*$ is strictly proper. Thus, together with $\langle W \rangle = 0$, equation (28) implies that ZA_1^{-1} is strictly proper. Thus, when (C.2) holds the optimal solution is also characterized by

$$ZA_1^{-1} \text{ strictly proper.}$$

Now, when $A^{-1} B$ is coprime (condition C.3) then $A_1^{-1} B_1 = B_1 A_1^{-1}$ is coprime. Thus, using Lemma A1 (see appendix) there exists a unique solution U , Z to equation (25) with ZA_1^{-1} strictly proper. This unique solution, as shown above, is the optimal solution.

Remark. Conditions (C.1) and (C.3) state that all reference and disturbance modes must also be present in the plant forward path. For unstable reference and disturbance models (those used to generate steps, ramps, shape-deterministic signals etc.) these conditions must already be satisfied when the optimal control problem has a solution (see Theorem 1). Thus, for a large and important class of problems these conditions will be satisfied.

Condition (C.2) also depends on the class of reference and disturbance signals of interest. It can normally be arranged that the transfer functions concerned are proper; this is certainly the case for the unstable models referred to above.

Note, finally, that whenever any of the conditions (C.1)–(C.3) do not hold all is not lost; we simply revert to the pairs of coupled polynomial matrix equations to calculate the controller.

5. Example

In this section we give an example which illustrates the results of Theorem 2. Consider a problem with the following data

$$\begin{aligned} B &= 2d + d^2 \quad C_1 \\ A &= 1 \quad 1.5d + 0.5d^2 \quad A_1 = 1 \\ C_2 &= A_1 \quad C_1 = C_2 = 1 \end{aligned}$$

$$\Omega = \Sigma = 1, \quad \phi_1 = \phi_2 = \phi_3 = \phi_4 = \phi_5 = 1.$$

It is easy to verify that condition (C.1) and conditions (C.2)–(C.3) of Theorem 2 are satisfied by this problem definition. Appealing to Theorem 2 we therefore expect the three implied equations alone to yield the optimal controller. To verify this we proceed, for each part of the controller, by first solving the implied equation. We then check the result by solving the full coupled equations for comparison.

Performing the spectral factorizations (7)–(9) we obtain:

$$\begin{aligned} D_1 &= 2.91 - 0.0811d + 0.172d^2 \\ D_2 &= 1.8 - 1.08d + 0.277d^2 \\ G &= 1.62 - 0.618d \end{aligned}$$

We now proceed to the controller calculations

Feedback controller. Solving the implied equation (21) for X and Y under the condition YA_d^{-1} strictly proper we obtain

$$X = 5.25 + 1.22d + 0.0953d^2, \quad Y = 1.68 - 0.678d.$$

For comparison we now solve the coupled equations (18) under the condition $\langle V \rangle = 0$ to obtain

$$\begin{aligned} X &= 5.25 + 1.22d + 0.0953d^2, \quad Y = 1.68 - 0.678d, \\ V &= -5.05d - 1.52d^2 \end{aligned}$$

and we observe that the implied equation has indeed yielded the unique optimal controller.

Reference controller. Solving the implied equation (23) under the condition MA_1^{-1} strictly proper we obtain

$$L = 4.71 + 0.778d + 0.106d^2, \quad M = 1.$$

Now solving the coupled equations (19) under the condition

$(N) = 0$ we obtain

$$L = 4.71 + 0.778d + 0.106d^2, \quad M = 1, \\ N = -4.15d - 1.45d^2.$$

Again, the solution of the implied equation under the condition of Theorem 2 has yielded the optimal controller.

Feedforward controller. Solving the implied equation (25) under the condition ZA_1^{-1} strictly proper we obtain

$$U = 2.91 + 1.14d, \quad Z = 1.57 - 0.571d.$$

Solving the couple of equations (20) under the condition $(W) = 0$ we obtain

$$U = 2.91 + 1.14d, \quad Z = 1.57 - 0.571d, \\ W = -3.32d - 0.73d^2.$$

This verifies that the implied equation has on its own yielded the optimal feedforward controller.

The results of this example are in agreement with Theorem 2; the conditions in the Theorem were satisfied by the particular problem data, and the three implied equations were therefore sufficient for the calculation of the optimal controller.

6. Conclusions

The design of optimal controllers for multivariable plants subject to measurable and unmeasurable disturbances has been considered. The optimal controller consists of feedback, reference tracking and measurable disturbance feedforward. In general, the optimal controller is determined by three pairs of coupled two-sided polynomial matrix equations. These coupled pairs can be reduced to three single one-sided equations, the implied polynomial matrix equations.

The conditions under which the implied polynomial matrix equations uniquely determine the optimal controller have been obtained in this paper.

References

- Grimble, M. J. (1984). Implicit and explicit LQG self-tuning controllers. *Automatica*, **20**, 661–669.
- Grimble, M. J. (1986). Feedback and feedforward LQG controller design. *Proc. Amer. Control Conf.*, Seattle.
- Grimble, M. J. (1988). Two-degrees of freedom feedback and feedforward optimal control of multivariable stochastic systems. *Automatica*, **24**, 809–817.
- Hunt, K. J. (1988). General polynomial solution to the optimal feedback/feedforward stochastic tracking problem. *Int. J. Control*, **48**, 1057–1073.
- Hunt, K. J. (1989). *Stochastic Optimal Control Theory with Application in Self-tuning Control*. Springer, Berlin.
- Hunt, K. J., M. J. Grimble, M. J. Chen and R. W. Jones (1986). Industrial LQG self-tuning controller design. *Proc. IEEE Conf. on Decision and Control*, Athens.
- Hunt, K. J., M. J. Grimble and R. W. Jones (1987a). LQG feedback and feedforward self-tuning control. *Proc. IFAC World Congress*, Munich.
- Hunt, K. J. and M. Šebek (1989a). Optimal multivariable regulation with disturbance measurement feedforward. *Int. J. Control*, **49**, 373–378.
- Hunt, K. J. and M. Šebek (1989b). Multivariable LQG self-tuning control with disturbance measurement feedforward. *Proc. IFAC Symp. on Adaptive Systems in Control and Signal Processing*, Glasgow.
- Hunt, K. J., M. Šebek and M. J. Grimble (1987b). Optimal multivariable LQG control using a single diophantine equation. *Int. J. Control*, **46**, 1445–1453.
- Kučera, V. (1979). *Discrete Linear Control*. Wiley, Chichester.
- Kwakernaak, H. and R. Sivan (1972). *Linear Optimal Control Systems*. Wiley, New York.
- Roberts, A. P. and M. M. Newmann (1988). Polynomial optimization of stochastic feedback control for stable plants. *IMA J. Math. Control Inform.*, **5**, 243–257.
- Šebek, M. (1983). Direct polynomial approach to discrete-time stochastic tracking. *Prob. Control Inform. Theory*, **12**, 293–302.
- Šebek, M., K. J. Hunt and M. J. Grimble (1988). LQG regulation with disturbance measurement feedforward. *Int. J. Control*, **47**, 1497–1505.
- Sternad, M. and T. Söderström (1988). LQG-optimal feedforward regulators. *Automatica*, **24**, 557–561.
- Youla, D. C., H. A. Jabr and J. J. Bongiorno (1976). Modern Wiener-Hopf design of optimal controllers—Part 2: the multivariable case. *IEEE Trans. Aut. Control*, **AC-21**, 319–338.

Appendix. Polynomial matrix solution

Lemma A1. Let M , N , P be given polynomial matrices (of suitable dimensions, M square, $\det M \neq 0$). Then the polynomial matrix equation

$$YN + XM = P$$

possesses a unique solution such that YM_2^{-1} is strictly proper, for a left coprime M_2 , N_2 given by

$$M_2^{-1}N_2 = NM^{-1}.$$

See Hunt *et al.* (1987b) for a proof of this result

Parameter Estimation Aspects in Adaptive Control*

F. GIRI,† J. M. DION,† L. DUGARD†§ and M. M'SAAD†

Key Words—Adaptive control; parameter estimation; convergence; stability.

Abstract—This paper provides a solution to the convergence of the parameter estimates to their true values in an ideal adaptive regulation context. The key design feature consists in the use of an asymptotically vanishing internally generated exciting sequence.

1. Introduction

In this paper, the following question is addressed: given a known order linear time invariant system, how does one design an adaptive regulator which ensures the convergence of both the parameter estimates error and the system output to zero?

The first objective (i.e. the convergence of the parameter estimates to their true values) can be obtained using an external persistently exciting signal as in Elliott *et al.* (1985), Goodwin and Teoh (1985) and Anderson and Johnstone (1985). However, the second objective (i.e. the asymptotic regulation of the system output to zero) cannot be obtained with such an approach. This drawback can be removed using an asymptotically vanishing probing signal. This issue has recently been investigated in Chen and Guo (1987) and Canudas de Wit (1987) using an external asymptotically vanishing sequence. Such an approach requires extra-prior knowledge about the system to be controlled.

In the present paper, we propose an answer to the above stated question. The main features of the underlying solution are: (1) It is carried out irrespective of both the parameter estimator and the control law; and (2) It involves an asymptotically vanishing exciting signal. Such a vanishing property is a consequence of the ideal framework we are concerned with. Of particular interest, the considered probing signal is internally generated, properly exploiting the parameter estimator properties.

Intuitively, the basic idea of this paper is then to build and to use an internally generated exciting signal that tends to zero sufficiently slowly to ensure the parameter convergence to the true parameters.

The paper is organized as follows. In Section 2 the control problem statement is given. The adaptive regulator is

described in Section 3. Section 4 is devoted to the main results of the paper.

2. Formulation of the control problem

2.1. The system to be controlled. Let us consider linear time-invariant discrete-time deterministic systems described by:

$$A(\theta^*, q^{-1})y(t) = B(\theta^*, q^{-1})u(t) \quad (2.1)$$

where $u(t)$ and $y(t)$ are the input and output, respectively, $A(\theta^*, q^{-1})$ and $B(\theta^*, q^{-1})$ are polynomials in the backward shift operator q^{-1} :

$$\begin{aligned} A(\theta^*, q^{-1}) &= 1 + \theta_1^* q^{-1} + \dots + \theta_n^* q^{-n} \\ B(\theta^*, q^{-1}) &= \theta_{n+1}^* q^{-1} + \dots + \theta_{n+m}^* q^{-n-m} \end{aligned} \quad (2.2)$$

and θ^* is a time-invariant vector in \mathcal{R}^{2n} :

$$\theta^* = [\theta_1^*, \dots, \theta_{2n}^*]^T \quad (2.3)$$

Further assumptions are made that

- A1: the system order n is known and
- A2: $|\det M_r(A(\theta^*, q^{-1}), B(\theta^*, q^{-1}))| \geq \varepsilon_n > 0$

where $M_r(X, Y)$ denotes the Sylvester matrix corresponding to polynomials X and Y .

The following equivalent representation of the system (2.1)–(2.3) will be convenient for identification purposes:

$$y(t) = \theta^{*T} \phi(t) \quad (2.4)$$

with

$$\phi(t) = [-y(t-1), \dots, -y(t-n), u(t-1), \dots, u(t-n)]^T \quad (2.5)$$

2.2. The basic control structure. We consider the following regulator structure:

$$R(\theta^*, q^{-1})u(t) + S(\theta^*, q^{-1})y(t) = 0 \quad (2.6)$$

with

$$\begin{aligned} R(\theta^*, q^{-1}) &= 1 + r_1(\theta^*)q^{-1} + \dots + r_{n-1}(\theta^*)q^{-(n-1)} \\ S(\theta^*, q^{-1}) &= s_0(\theta^*) + s_1(\theta^*)q^{-1} + \dots + s_{n-1}(\theta^*)q^{-(n-1)} \end{aligned} \quad (2.7)$$

The $R(\theta^*, q^{-1})$ and $S(\theta^*, q^{-1})$ polynomials are determined according to a given control design technique (e.g. model reference, pole placement, linear quadratic, long range predictive, ...).

More precisely, the correspondence $K(\cdot)$ between the system parameter θ^* and the regulator parameter:

$$K(\theta^*) = [r_1(\theta^*) \dots r_{n-1}(\theta^*) \ s_0(\theta^*) \dots s_{n-1}(\theta^*)]^T \quad (2.8)$$

is continuous. For stability purposes, we consider the following definition.

Definition 1. Given a control evaluation function $K(\cdot)$, a subset D_ρ of \mathcal{R}^{2n} is said to be admissible with respect to the function $K(\cdot)$ if there exists a scalar $\rho: 0 \leq \rho < 1$ such that for

* Received 13 July 1987; revised 6 April 1988; revised 26 October 1989; received in final form 12 July 1990. The original version of this paper was presented at the IFAC Symposium on Identification and Systems Parameter Estimation which was held in Beijing, People's Republic of China during August, 1988. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor A. Bagchi under the direction of Editor P. C. Parks.

† Laboratoire d'Automatique de Grenoble, ENSIEG, BP 46, 38402 Saint-Martin-d'Hères, France.

‡ Present address: LA21, Ecole Mohammadia d'Ingénieurs, B.P. 765, Rabat-Agdal, Morocco.

The authors are all also with the Greco C.N.R.S. "Adaptive Systems".

§ Author to whom all correspondence should be addressed.

all $\theta \in D_\theta$,

$$P(\theta, q^{-1}) \triangleq A(\theta, q^{-1})R(\theta, q^{-1}) + B(\theta, q^{-1})S(\theta, q^{-1}) \\ = 0, \quad \Rightarrow |q| \approx \rho < 1 \quad (2.9)$$

and if there exists a positive constant K_θ such that for all $\theta_1, \theta_2 \in D_\theta$,

$$\|K(\theta_1) - K(\theta_2)\| \leq K_\theta \|\theta_1 - \theta_2\|. \quad \nabla$$

The implication (2.9) means that the regulator (2.6)–(2.7) stabilizes the class of systems (2.1)–(2.3). We will make the further assumption:

A3: θ^* is an interior point of D_θ .

It is worth noticing that a complete knowledge of D_θ is not necessary. It is sufficient to check that a given θ belongs to D_θ , using any stability criterion.

For example, when the pole placement (respectively L.Q.) control is used, it is sufficient to check that the model is sufficiently controllable (respectively stabilizable).

The problem of concern is to design an indirect adaptive regulator which ensures the exponential convergence of the input-output signals to zero.

3. The indirect adaptive regulation algorithm

The considered indirect adaptive regulator is simply obtained by combining a parameter estimator with the linear control law (2.6)–(2.7). The involved parameter estimator and the adaptive control law are given below.

3.1. The parameter estimation algorithm. Instead of explicitly giving a particular identification algorithm providing the estimates $\hat{\theta}(t)$ of θ^* , we state three conditions to be satisfied by the algorithm. By doing so, it will be clear that these conditions are the only ones which are crucial for the closed loop global convergence. Such an approach has been followed in previous studies (Samson, 1982; De Larminat, 1981). The involved conditions are the following:

$$C1 \quad \lim_{t \rightarrow \infty} \phi(t)^T \hat{\theta}(t) = 0; \quad \theta \neq \hat{\theta}(t) = \theta^*$$

$$C2 \quad \lim_{t \rightarrow \infty} \|\hat{\theta}(t) - \hat{\theta}(t-1)\| = 0$$

$$C3 \quad \|\hat{\theta}(t) - \theta^*\| \text{ converges to some value}$$

The above conditions C1–C3 are straightforwardly interpreted. Roughly speaking, condition C1 means that no difference can be asymptotically observed between the system and its estimated model. C2 states that the estimated model is more and more slowly time-varying. C3 insures the uniform boundedness of the parameter estimates. Conditions C1–C3 are satisfied by available identification algorithms, e.g. projection algorithm, least squares, L.S with covariance resetting (Goodwin and Sin, 1984).

3.2. The adaptive control law. For any time instant t , let k be the unique integer satisfying: $4nk \leq t \leq 4n(k+1) - 1$ and $\theta_k(t)$ be the parameter sequence defined by:

$$\theta_k(t) = K(\hat{\theta}(t)) \quad (3.1a)$$

where

$$\tau = \max \{i/t \leq 4nk \text{ and } \hat{\theta}(i) \in D_\theta\} \quad (3.1b)$$

The adaptive law considered in this paper, is given by:

$$R(\theta_k(t), q^{-1})u(t) + S(\theta_k(t), q^{-1})y(t) = \beta(t)m(t) \quad (3.2a)$$

with

$$m(t) = \sigma m(t-1) + \max \{\|\phi(t)\|, 1\}; \quad 0 < \sigma \leq 1,$$

$$m(0) > 0 \quad (3.2b)$$

$$\beta(t) = \begin{cases} \gamma(t) & \text{if } t = 4nk + 2n - 1 \\ 0 & \text{otherwise} \end{cases} \quad (3.2c)$$

where

$$\gamma(t) = \left\{ \sum_{i=\tau}^t |\bar{\phi}(t+i)^T \hat{\theta}(t+i)| \right. \\ \left. + \sum_{i=4nk+1}^t \|\hat{\theta}(t+i) - \hat{\theta}(t+i-1)\| \right\} \quad (3.2d)$$

$$\bar{\phi}(t) \triangleq \phi(t)/m(t). \quad (3.2e)$$

Notice that the above control law is implementable since, $\gamma(t)$, $\beta(t)$ and $m(t)$ entirely depend on information that are available at time t .

The main features of the above adaptive control law are the following: (1) the controller parameter sequence $\theta_k(t)$ is frozen over an horizon of $4n$ sampling periods as pointed out by equations (3.1a) and (3.1b) and is constructively admissible; and (2) an internal impulse exciting sequence is periodically added to ensure the parameter convergence to their true values as it will be shown. Due to conditions C1–C3, the sequence $\{\beta(t)\}$ is asymptotically vanishing infinitely more slowly than the parameter estimates as well as the estimation error convergence.

Though the considered exciting sequence is similar to that proposed in Kreisselmeier and Smith (1986) or Polderman (1987), it differs from the definition of $\{\beta(t)\}$. Such a difference gives rise to the potential result of this paper, namely the convergence of the parameter estimates to their true values, via a vanishing exciting signal.

4. Closed loop global stability and convergence

This section is devoted to the main result of the paper, namely:

Theorem 1. Consider the system (2.1)–(2.3), subject to assumptions A1–A3, in closed loop with the adaptive control law (3.1)–(3.2). Then for any bounded arbitrary initial condition $\theta(0)$, $u(0)$ and $y(0)$ one has:

- (i) $\lim_{t \rightarrow \infty} \hat{\theta}(t) = \theta^*$
- (ii) the input-output signals are exponentially vanishing to zero. ∇

The main ingredients to establish this result are the following properties of the closed loop signals.

Proposition 1. There exists a positive constant K_m such that for all t , one has:

$$m(t+1) \geq K_m m(t) \quad \nabla$$

The proof is straightforward (see Giri *et al.*, 1987a).

Proposition 2. There exists a positive scalar δ , such that for all $j \in \mathcal{N}$, and for all unit vector $w \in \mathcal{R}^{2n}$ one has:

$$|w^T \bar{\phi}(4nj + \tau)| \geq \delta \gamma(4nj + 2n - 1)$$

for at least one $\tau \in [-4n + 1, -1]$.

The proof of this "richness" property is given in Appendix A.

Proof of Theorem 1.

(i) Let $t, k, \tau \in \mathcal{N}$ such that: $4nk \leq t \leq 4n(k+1) - 1$ and $-8n + 2 \leq \tau \leq -1$. One has:

$$|\bar{\phi}(t + \tau)^T \hat{\theta}(t)| \leq |\bar{\phi}(t + \tau)^T \hat{\theta}(t + \tau)| \\ + |\bar{\phi}(t + \tau)^T (\hat{\theta}(t + \tau) - \hat{\theta}(t))| \\ \leq \sum_{i=\tau}^t |\bar{\phi}(t+i)^T \hat{\theta}(t+i)| \\ + \sum_{i=4nk+1}^t \|\hat{\theta}(t+i) - \hat{\theta}(t+i-1)\| \quad (4.1)$$

where the second inequality follows using the fact that $\|\bar{\phi}(t)\| \leq 1$, for all $t \in \mathcal{N}$. Otherwise, using Proposition 2, it follows that for at least one τ' : $-4n + 1 \leq \tau' \leq -1$, one has:

$$|\bar{\phi}(4nk + \tau')^T \hat{\theta}(t)| \geq \delta \gamma(4nk + 2n - 1) \|\hat{\theta}(t)\|$$

which yields:

$$|\bar{\phi}(t + (4nk - t) + \tau')^T \hat{\theta}(t)| \geq \delta \gamma(4nk + 2n - 1) \|\hat{\theta}(t)\|.$$

Hence, for all $t, k \in \mathcal{N}$, such that: $4nk \leq t \leq 4n(k+1) - 1$, there exists a τ' : $-8n + 2 \leq \tau' \leq -1$, so that:

$$|\bar{\phi}(t + \tau) \hat{\theta}(t)| \geq \delta \gamma(4nk + 2n - 1) \|\hat{\theta}(t)\|. \quad (4.2)$$

Combining (4.1) and (4.2) yields: for all $t, k \in \mathcal{N}$, such

that: $4nk \leq t \leq 4n(k+1) - 1$.

$$\begin{aligned} \|\tilde{\theta}(t)\| &\leq \left\{ \sum_{i=kn+2}^{t-1} |\tilde{\phi}(t+i)^T \tilde{\theta}(t+i)| + \sum_{i=kn+1}^n \|\tilde{\theta}(t+i) \right. \\ &\quad \left. - \tilde{\theta}(t+i-1)\| \right\} / \delta\gamma(4nk - 2n - 1) \\ &\leq [\gamma(t)]^2 / \delta\gamma(4nk - 2n - 1) \end{aligned} \quad (4.3)$$

where the second inequality follows from (3.2d). Using conditions C1-C3 it follows that $\gamma(t) \rightarrow 0$ as $t \rightarrow \infty$, which, together with (4.3), leads to:

$$\lim_{t \rightarrow \infty} \|\tilde{\theta}(t)\| = 0.$$

This completes the proof. \square

(ii) The second part of the theorem is by now standard (Goodwin and Sin, 1984) since the sequences $\{\hat{\theta}(t)\}$ and $\{\beta(t)\}$ converge to θ^* and 0, respectively [see Giri *et al.* (1987b) for more details].

Remark. In the nonideal case (unmodeled dynamics and bounded disturbances) the above results are no longer valid.

5. Concluding remarks

Parameter estimation aspects have been investigated in this paper. An asymptotically vanishing internally generated exciting sequence has been periodically added to the control signal, ensuring the asymptotic parameter error convergence to zero and henceforth the estimated model admissibility. Furthermore, the asymptotic regulation of the system output to zero is achieved.

References

- Anderson, B. D. O. and R. M. Johnstone, (1985). Global adaptive pole positioning *IEEE Trans. Aut. Control*, **AC-30**, 11-22.
- Canudas de Wit, C. (1987). Adaptive control for partially known systems. Ph.D. Thesis, INP Grenoble.
- Chen, M. F. and L. Guo (1987). Asymptotically optimal adaptive control with consistent parameter estimates. *SIAM J. Control Optimiz.*, **25**, 558-575.
- De Larminat, Ph. (1981). Unconditional stabilization of linear discrete systems via adaptive control. *Syst. Control Lett.*, **1**, 7-11.
- Elliott, H., R. Cristi and M. Das (1985). Global stability of adaptive pole placement algorithms. *IEEE Trans. Aut. Control*, **AC-30**, 348-356.
- Giri, F., J. M. Dion, M. M'Saad and L. Dugard (1987a). A globally convergent pole placement indirect adaptive controller. *Proc. 26th IEEE CDC* Los Angeles, CA, **1**, 372-377.
- Giri, F., J. M. Dion, L. Dugard and M. M'Saad (1987b). Parameter estimation aspects in adaptive control. *Proc. IFAC Symp. on Identification and Systems Parameter Estimation*, Beijing, China.
- Goodwin, G. C. and K. S. Sin (1984). *Adaptive Filtering Prediction and Control*. Prentice Hall, Englewood Cliffs, New Jersey.
- Goodwin, G. C. and E. K. Teoh (1985). Persistency of excitation in the presence of possibly unbounded signals. *IEEE Trans. Aut. Control*, **AC-30**, 595-597.
- Kreisselmeier, G. and M. Smith (1986). Stable adaptive regulation of arbitrary n th-order plants. *IEEE Trans. Aut. Control*, **AC-31**, 299-305.
- Polderman, J. W. (1987). A state approach to the problem of adaptive pole assignment. Report OS-R8704, CMCS, CWI, Amsterdam, Netherlands.
- Samson, C. (1982). An adaptive LQ controller for nonminimum phase systems. *Int. J. Control*, **35**, 1-28.

Appendix A: Proof of Proposition 2

The proof is done in three steps. In step 1, a new closed loop state vector $x(t)$ is defined and related to $\phi(t)$. Proposition 2 is shown to hold for $x(t)$, in step 2. It then follows readily from steps 1 and 2.

Step 1. Following Kreisselmeier and Smith (1986) the plant (2.1) can be described by the following controllable representation.

$$\begin{aligned} A(\theta^*, q^{-1})\xi(t) &= u(t) \\ y(t) &= B(\theta^*, q^{-1})\xi(t) \end{aligned} \quad (a1)$$

where $\xi(t)$ is an internal state variable. The substitution of (a1) in (3.2a) yields

$$\begin{aligned} [R(\theta, (t), q^{-1})A(\theta^*, q^{-1}) \\ + S(\theta, (t), q^{-1})B(\theta^*, q^{-1})]\xi(t) &= f(t) \end{aligned} \quad (a2)$$

where $f(t) = \beta(t)m(t)$. Defining the state vector $x(t)^T = [\xi(t-1) \dots \xi(t-2n)]$, equation (a2) can be equivalently written,

$$x(t+1) = G(t)x(t) + gf(t) \quad (a3)$$

where

$$G(t) = \begin{bmatrix} -\gamma_1(t) & \dots & -\gamma_{2n-1}(t) & 0 \\ 1 & 0 & & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & & 1 & 0 \end{bmatrix}, \quad g = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (a4)$$

and $\gamma_i(t)$ is the coefficient of q^{-i} in the left-hand side of (a2).

Substituting (a1) in (2.5) yields

$$\begin{aligned} \phi(t)^T &= [-B(\theta^*, q^{-1}), \dots, -q^{-(n-1)}B(\theta^*, q^{-1}), \\ &\quad A(\theta^*, q^{-1}), \dots, q^{-(n-1)}A(\theta^*, q^{-1})]\xi(t-1) \end{aligned} \quad (a5)$$

Since $A(\theta^*, q^{-1})$ and $B(\theta^*, q^{-1})$ are coprime, the $2n$ -polynomials on the right-hand side of (a5) are linearly independent over the reals. As $x(t)^T = [1 q^{-1} \dots q^{-(2n-1)}]\xi(t-1)$, it follows from (a5) that there exists a full rank $2n \times 2n$ matrix H such that

$$\phi(t) = Hx(t). \quad (a6)$$

Step 2. Following Kreisselmeier and Smith (1986), one obtains from (a3), after some computations

$$\begin{aligned} &\sum_{i=0}^{2n-1} \gamma_{2n-1-i}(t+2n-1)x(t+k+i+2n) \\ &= M(t+2n-1)F(t+k+2n-1) \\ &\quad + \sum_{i=0}^{2n-1} \gamma_{2n-1-i}(t+2n-1) \\ &\quad \times \left\{ \sum_{j=0}^i G(t+2n-1)^j [G(t+k+j+2n-1) \right. \\ &\quad \left. - G(t+2n-1)]x(t+k+j+2n-1) \right\} \end{aligned} \quad (a7)$$

where $\gamma_0(t) = 1$, $-2n+1 \leq k \leq 0$ and for all t ,

$$M(t) = [g, G(t)g, \dots, G(t)^{2n-1}g] \begin{bmatrix} \gamma_{2n-1}(t) & \dots & \gamma_1(t) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma_1(t) & \dots & 1 & 0 \\ 1 & & & \end{bmatrix} \quad (a8)$$

$$F(t)^T = [f(t), \dots, f(t+2n-1)]. \quad (a9)$$

Let $t = 4nk'$ for some $k' \in \mathcal{N}$. From (3.1a, b) it follows that the controller parameters $r_i(t)$, $s_i(t)$ ($0 \leq i \leq n-1$) are "frozen" over the interval $[t, t+4n-1]$. This implies that for all $\tau: t+1 \leq \tau \leq t+4n-1$, $\gamma_i(\tau) = \gamma_i(t-1)$ and therefore, using (a4), $G(\tau) = G(t-1)$. Hence, for all

$k: -2n+1 \leq k \leq 0$ and $k' \in \mathcal{N}$, we have

$$\left\| \sum_{i=0}^{2n-1} \gamma_{2n-i-1}(4nk'+2n-1) \left\{ \sum_{j=0}^i G(4nk'+2n-1)^{j-1} \cdot [G(4nk'+k+j+2n-1) - G(4nk'+2n-1)] \cdot x(4nk'+k+j+2n-2) \right\} \right\| = 0. \quad (\text{a10})$$

Let v be an arbitrary unit $2n$ -vector. Premultiplication of (a7) by v^T and use of (a10), yields, for all $k: -2n+1 \leq k \leq 0$ and all $k' \in \mathcal{N}$,

$$\left| \sum_{i=0}^{2n-1} \gamma_{2n-i-1}(4nk'+2n-1) v^T x(4nk'+k+i+2n) \right| \leq |v^T M(4nk'+2n-1) F(4nk'+k+2n-1)|. \quad (\text{a11})$$

Using assumption A5 it follows from (3.1a, b) that the controller parameters are uniformly bounded. Then, from (a2) there exists a positive constant K_γ such that for all $i: |\gamma_i(t)| \leq K_\gamma (0 \leq i \leq 2n-1)$. This together with (3.2b) yields for all $k: -2n+1 \leq k \leq 0$ and $k' \in \mathcal{N}$:

$$\begin{aligned} & \left| \sum_{i=0}^{2n-1} \gamma_{2n-i-1}(4nk'+2n-1) v^T x(4nk'+k+i+2n) \right| \\ & \leq \sum_{i=0}^{2n-1} K_\gamma m(4nk'+k+i+2n) |v^T \tilde{x}(4nk'+k+i+2n)| \\ & \leq 2nK_\gamma K_1 m(4nk'+4n-1) \max_{1 \leq i \leq 4n-1} |v^T \tilde{x}(4nk'+\tau)| \end{aligned} \quad (\text{a12})$$

using Proposition 1 for some positive constant K_1 , where $\tilde{x} = x/m$.

From (a8), $M(t)$ is an invertible triangular matrix with positive, bounded away from zero, singular values for any t . Let σ_M a positive lower bound of these singular values.

$$\begin{aligned} & F(4nk'+k+2n-1) \\ & = [0 \cdots 0 \ 1 \ 0 \cdots 0]^T \{ \gamma(4nk'+2n-1) m(4nk'+2n-1) \} \end{aligned}$$

where 1 is at the $(1-k)$ th entry and $f(t)$ is replaced by $\beta(t)m(t)$. Thus for at least one $k: -2n+1 \leq k \leq 0$ and all $k' \in \mathcal{N}$:

$$\begin{aligned} & |v^T M(4nk'+2n-1) F(4nk'+k+2n-1)| \\ & \geq \frac{\sigma_M \gamma(4nk'+2n-1) m(4nk'+2n-1)}{\sqrt{2n}}. \end{aligned} \quad (\text{a13})$$

Combining (a11)–(a13) it follows that for all $k' \in \mathcal{N}$:

$$\begin{aligned} & \max_{1 \leq i \leq 4n-1} |v^T \tilde{x}(4nk'+\tau)| \\ & \geq \frac{\sigma_M \gamma(4nk'+2n-1) m(4nk'+2n-1)}{2n\sqrt{2n}K_1K_\gamma} \frac{1}{m(4nk'+4n-1)}. \end{aligned} \quad (\text{a14})$$

Using Proposition 1 and inequality (a14) there exists a positive constant K_2 such that, for all $k' \in \mathcal{N}$:

$$\max_{1 \leq i \leq 4n-1} |v^T \tilde{x}(4nk'+\tau)| \geq \frac{\sigma_M \gamma(4nk'+2n-1)}{2n\sqrt{2n}K_2K_1K_\gamma}. \quad (\text{a15})$$

Step 3. Now let ω be an arbitrary unit $2n$ -vector and define the unit $2n$ -vector $v^T = \omega^T H / \|\omega^T H\|$. Using (a6) one obtains for all t and τ

$$|\omega^T \tilde{\phi}(t+\tau)| = |\omega^T H \tilde{x}(t+\tau)| \leq \sigma_{\min}(H) |v^T \tilde{x}(t+\tau)| \quad (\text{a16})$$

where $\sigma_{\min}(H) > 0$ is the minimum singular value of H . Letting $\delta = \sigma_M [\sigma_{\min}(H)] / [2n\sqrt{2n}K_2K_1K_\gamma]$, (a15) and (a16) imply for all $k' \in \mathcal{N}$

$$\max_{1 \leq i \leq 4n-1} |\omega^T \tilde{\phi}(4nk'+\tau)| \geq \delta \gamma(4nk'+2n-1). \quad (\text{a17})$$

Letting $k' = j-1$ it follows from (a17) that for all $j \in \mathcal{N}$:

$$\max_{1 \leq i \leq 4n-1} |\omega^T \tilde{\phi}(4nj+\tau)| \geq \delta \gamma(4nj-2n-1) \quad (\text{a18})$$

which establishes Proposition 2. $\nabla \nabla \nabla$

Brief Paper

Estimation Theory for Nonlinear Models and Set Membership Uncertainty*

M. MILANESE† and A. VICINO‡

Key Words—Estimation theory; set membership uncertainty; unknown but bounded noise; uncertainty intervals; nonlinear model

Abstract—In this paper we study the problem of estimating a given function of a vector of unknowns, called the problem element, by using measurements depending nonlinearly on the problem element and affected by unknown but bounded noise. Assuming that both the solution sought and the measurements depend polynomially on the unknown problem element, a method is given to compute the axis-aligned box of minimal volume containing the feasible solution set, i.e. the set of all unknowns consistent with the actual measurements and the given bound on the noise. The center of this box is a point estimate of the solution, enjoying useful optimality properties. The sides of the box represent the intervals of possible variation of the estimates. It is shown how important problems, like parameter estimation of exponential models, time series prediction with ARMA models and parameter estimation of discrete time state space models, can be formalized and solved by using the developed theory.

1. Introduction

IN THIS paper the following problem, referred to as the (generalized) estimation problem, is addressed. Consider a problem element λ , for example the vector of parameters of a dynamic system or a time function. We are interested in evaluating a vector valued function $S(\lambda)$ of this problem element (for example, some functions of parameters of the dynamic system or particular values of the time function). The element λ is not exactly known and we have only some information on it. In particular, we assume that it belongs to a set K of possible problem elements and that information on λ is given by the knowledge of a function $F(\lambda)$, representing measurements performed on variables depending on λ . We suppose that exact measurements are not available and actual measurements y are corrupted by some error ρ according to the equation

$$y = F(\lambda) + \rho \quad (1)$$

The estimation problem consists in finding an algorithm (estimator) ϕ providing an approximation $\phi(y) \approx S(\lambda)$ as a function of the available data y and in evaluating a measure of the approximation error.

Many different problems such as linear and nonlinear regressions, parameter or state estimation of dynamic systems, state space and ARMA models prediction, filtering, smoothing, time series forecasting, interpolation, function approximation can be formulated in a general unifying framework based on the above concepts.

* Received 3 November 1988; revised 30 May 1989; revised 13 March 1990; received in final form 5 June 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. Lozano-Leal under the direction of Editor P. C. Parks.

† Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.

‡ Dipartimento di Sistemi e Informatica, Università di Firenze, Via di Santa Marta 3, 50139 Firenze, Italy. Author to whom all correspondence should be addressed.

The solution of the estimation problem depends on the type of assumptions made on ρ . Most of the investigated cases in the literature on estimation theory are undoubtedly related to the assumption that the error vector ρ is statistically modelled as an at least partially known probability distribution. Within this context, the most important and widely used results are related to the theory of maximum likelihood estimators (MLE). Despite the large amount of theoretical results developed on MLE, the application to real world problems may not be appropriate due to a number of possible drawbacks. These include:

1. Actual computation of MLE usually requires a search of the global extremum of functions which are in general multimodal. Since general optimization algorithms (including the so called global ones based on random search) are not guaranteed to achieve the global extremum, the estimate obtained may be far from being an MLE;
2. Even though MLE are asymptotically efficient, it is difficult to evaluate whether the available data are sufficient to ensure that the covariance matrix estimate is "close" to the Cramer-Rao lower bound;
3. For small data sets, it is useful to have lower and upper bounds of the estimate covariance matrix; indeed, tight upper bounds are difficult to evaluate. Moreover, in this condition even the evaluation of the Cramer-Rao lower bound may be not significant;
4. It is difficult to evaluate the effect of nonexact matching of the assumed statistical hypotheses on ρ . In particular, there is no theory for taking into account the presence in ρ of modelling errors.

In more recent years a new approach, referred to as "set membership error description" or "unknown but bounded error (UBBE)", has been investigated [see Milanese, (1989) for a survey on the topic]. In this case, the error vector ρ is assumed to be an element of an admissible error set described by a norm operator as

$$\|\rho\| \leq \epsilon \quad (2)$$

where ϵ is a known quantity. A case of great concern is when l_1 norms are adopted; in this case, each component of the error vector is known to be bounded by given values. Motivation for this kind of error representation is the fact that in many practical cases the UBBE information is more realistic than statistical assumptions with respect to the measurement error (Schweppe, 1973; Milanese and Belforte, 1982). In this context, a possible approach to the estimation problem consists in finding the set of values $S(\lambda)$ (feasible solution set) such that λ is consistent with the measurements y and the error model (2). Any element of this set represents a possible estimate, although the center or the minimum norm element of the set enjoy interesting optimality properties (Traub and Woźniakowski, 1980; Micchelli and Rivlin, 1977; Kaciewicz *et al.*, 1986; Milanese *et al.*, 1986). The size of the set represents a measure of the estimate reliability.

Unfortunately, an exact representation of the feasible solution set is in general not simple, since it may not be

convex and not even connected. It is therefore convenient to look for simpler, although approximate, descriptions of this set. To this extent, the use of simply shaped sets, like axis-aligned boxes (referred to as boxes for short) or ellipsoids, has been proposed to approximate the feasible solution set (Milanese and Belforte, 1982; Fogel and Huang, 1982). Ellipsoids may approximate the shape of the feasible solution set better than boxes. Unfortunately, algorithms for computing ellipsoidal approximations are known for linear $S(\cdot)$ and $F(\cdot)$ only (Fogel and Huang, 1982; Norton, 1987). Moreover, the obtained approximations may not be tight (Belforte and Bona, 1985; Norton, 1987). On the other hand, important information can be obtained by box approximation. In particular, the minimal volume box containing the feasible solution set (MOB, minimal outer box) has the following properties

- The length of each of its sides along the corresponding i th coordinate axis gives the maximum range of possible variation of $S(\lambda_i)$, (called uncertainty interval UI_i).
- The center of MOB is the (Chebyshev) center of the feasible solution set and hence it is an estimate of $S(\lambda)$ enjoying several optimality properties (Milanese *et al.*, 1986; Kaciewicz *et al.*, 1986).

For linear problems, the MOB can be easily computed by solving suitable linear programming problems (Milanese and Belforte, 1982). Unfortunately, many practical estimation problems, even if related to linear dynamic models, lead to nonlinear $S(\cdot)$ and $F(\cdot)$ (see Section 3). Several approaches have been proposed to evaluate MOB when $F(\cdot)$ is nonlinear and $S(\cdot)$ is identity. In Clement and Gentil (1988) a solution is found in the case in which (1) represents model output-error equations. In Belforte and Milanese (1981) a method of successive linearization is proposed to construct a sequence of boxes contained in the MOB, but no guarantee of convergence to the MOB is given. In Smit (1983) and Walter and Piet-Lahanier (1986) optimization methods are used to construct the boundary of the feasible solution set. In particular, the random search algorithm used in Walter and Piet-Lahanier (1986) generates a sequence of boxes contained in the MOB and converging monotonically to it with probability one. However, this convergence property may not be useful in practice, because no estimate is given of the distance of the achieved solution from the global solution.

In this paper we show that if $S(\cdot)$ and $F(\cdot)$ are polynomial functions, a sequence of boxes contained in the MOB can be constructed, converging to it. Moreover, an estimate of the distance of the estimated box from the MOB is provided at each iteration. It is also shown that the hypothesis of $S(\lambda)$ and $F(\lambda)$ polynomial covers large classes of problems of practical interest such as, for example, the identification of multieponential, ARMA and state space discrete time models.

The paper is organized as follows. Section 2 introduces the spaces and operators needed to build a general framework for estimation problems. In Section 3 it is shown how some significant estimation problems lead to polynomial $S(\lambda)$ and $F(\lambda)$. Section 4 presents an optimization algorithm which allows one to derive a guaranteed global solution for the class of polynomial problems mentioned above. The effectiveness of the proposed approach is demonstrated by some examples reported in Section 5.

2. A general framework for estimation problems

Let Λ be a linear normed n -dimensional space on the real field (called the *problem element space*). Consider a given operator S , called the *solution operator* mapping Λ into Z .

$$S: \Lambda \rightarrow Z \quad (3)$$

where Z is a linear normed l -dimensional space on the real field. In estimation theory, the aim is to estimate an element $S(\lambda)$ belonging to the *solution space* Z , knowing approximate information about the element λ .

The available information on the problem is contained in the space Λ and in an additional linear space Y which will be introduced below. The first kind of information, which is

referred to as *a priori* information, generally consists in assuming that λ belongs to a subset K of Λ . In our development, we will deal with problems for which either $K = \Lambda$ (i.e. no *a priori* information is available), or K is given as

$$K = \{\lambda \in \Lambda : \|P(\lambda - \lambda_0)\| \leq 1\} \quad (4)$$

where P is a linear operator and λ_0 is a known problem element. Despite the above assumption, many of the results presented in the paper hold also for more general structures of the set K . Concerning the second kind of information, we assume that some function $F(\lambda)$ is given; F , called *information operator*, is a map from Λ to a linear normed m -dimensional space Y (called *measurement space*)

$$F: \Lambda \rightarrow Y. \quad (5)$$

We assume that Z and Y are equipped with (weighted) l_2 norms \dagger .

In general, due to the presence of noise, exact information $F(\lambda)$ about λ is not available and only perturbed information y is given. In this context, information uncertainty ρ is assumed to be additive, i.e.

$$y = F(\lambda) + \rho \quad (6)$$

where the error term ρ is unknown but bounded by a given positive value ϵ according to an l_2^* norm

$$\|\rho\|_{l_2^*} \leq \epsilon. \quad (7)$$

Notice that the use of an l_2^* norm in the measurement space Y allows us to consider different error bounds on every measurement. An *algorithm* ϕ is an operator (in general nonlinear) from Y into Z

$$\phi: Y \rightarrow Z \quad (8)$$

which provides an approximation $\phi(y)$ of $S(\lambda)$ using the available data y . Such an algorithm is also referred to as an *estimator*.

As a simple example of how a specific estimation problem fits into the general framework outlined above, consider the problem of parameter estimation of a time function belonging to a finite dimensional space, using data obtained by sampling and measuring it at a number of instants. Roughly speaking, the problem element space is the space of the considered class of functions, identified as the space of the unknown function parameters; the space Y is the space of available samples (possibly corrupted by noise); the solution operator is the identity operator and the information operator is the sampling operator.

Now, we introduce the following set, which plays a key role in the development of the theory

$$T(y) = \{\lambda \in K : \|y - F(\lambda)\|_{l_2^*} \leq \epsilon\} \quad (9)$$

The set $T(y)$ contains all λ compatible with the information F , the data y and the bound ϵ on the noise; $S(T(y))$ represents the already mentioned *feasible solution set*. We make some technical assumptions about this set. First, there exists a set $Y_0 \subseteq Y$ such that for each $y \in Y_0$, $T(y)$ is nonempty, i.e. the model structure is able to represent all the data y belonging to the set Y_0 . Secondly, $T(y)$ does not contain isolated (discrete) points. Third, $T(y)$ is bounded; if this were not true, $F(\lambda)$ would be too poor to solve the problem with finite error, indicating the presence of unidentifiability conditions in the problem formulation. Notice that the above hypotheses are almost always implicitly assumed in the great majority of identification problems.

Algorithm approximation will be measured according to the following *local* and *global* errors:

1. *Y-local* error $E(\phi, y)$

$$E(\phi, y) = \sup_{\lambda \in T(y)} \|S(\lambda) - \phi(y)\| \quad (10)$$

2. *Λ -local* error $E(\phi, \lambda)$

$$E(\phi, \lambda) = \sup_{y: \|y - F(\lambda)\|_{l_2^*} \leq \epsilon} \|S(\lambda) - \phi(y)\| \quad (11)$$

- \dagger A weighted l_2 norm, denoted by l_2^* , is defined as

$$\|y\|_{l_2^*} = \max_i w_i |y_i|, \quad w_i > 0.$$

3. global error $E(\phi)$

$$E(\phi) = \sup_{y \in Y_0} E(\phi, y) = \sup_{\lambda \in \Lambda} E(\phi, \lambda). \quad (12)$$

Algorithms minimizing these types of errors are called *Y-locally*, *Λ -locally* and *globally* optimal, respectively. Notice that the above errors, and related optimality concepts, are relevant to estimation problems. In fact, the Λ -local error measures the maximum uncertainty of the estimates induced by the perturbation affecting the exact information $F(\lambda)$, for a given problem element λ . On the other hand, the *Y*-local error measures the uncertainty affecting an estimate of $S(\lambda)$, for a given set of data y , λ being unknown. The global error represents a *worst case* cost function, in the sense that it measures the largest estimation uncertainty arising from the worst data realization and the worst problem element in the set $T(y)$ of admissible problem elements.

As already mentioned, the set $T(y)$ plays a key role in the present theory. In particular, if $z^* \in Z$ is the Chebyshev center of $S(T(y))$,† the algorithm ϕ^* , called the *central algorithm*, defined by

$$\phi^*(y) = z^* \quad (13)$$

is known to be *Y-locally* and *globally* optimal (Traub and Woźniakowski, 1980; Micchelli and Rivlin, 1977). In addition, Λ -local optimality of ϕ^* has been proven under mild assumptions in Milanese *et al.* (1986), Kaciewicz *et al.* (1986) for the case where $S(\cdot)$ and $F(\cdot)$ are linear.

Important information can be also derived from the knowledge of the quantities z_i^m and z_i^M , solutions of the following optimization problems

$$z_i^m = \inf_{\lambda \in T(y)} (S(\lambda))_i; \quad i = 1, \dots, l \quad (14)$$

$$z_i^M = \sup_{\lambda \in T(y)} (S(\lambda))_i; \quad i = 1, \dots, l$$

More precisely, we observe that the intervals

$$UI_i = [z_i^m, z_i^M], \quad i = 1, \dots, l, \quad (15)$$

represent the range of possible variations of the unknown solution components. The MOB containing $S(T(y))$ is obtained as the cartesian product of the UI_i

$$MOB = [UI_1 \times UI_2 \times \dots \times UI_l]. \quad (16)$$

The central algorithm ϕ^* can be computed componentwise, as Milanese and Tempo (1985)

$$(\phi^*(y))_i = z_i^* = (z_i^m + z_i^M)/2; \quad i = 1, \dots, l \quad (17)$$

Unfortunately, finding global solutions of problems (14) is in general a difficult task. If no further assumptions on S and F are made, the use of general global optimization algorithms assures at most convergence in probability to global extrema. More importantly, these methods do not provide any measure of how far is the computed solution from the global minimum (see e.g. Pardalos and Rosen, 1987; van Laarhoven and Aarts, 1987). However, in many estimation problems $S(\lambda)$ and $F(\lambda)$ are polynomial functions of λ (as shown in the next section). In these cases, it is possible to design algorithms (as the one presented in Section 4) which ensure certain convergence to global extrema, and give at each step a measure of how far is the actual solution from the global one.

3. Nonlinear estimation of dynamic models

As already mentioned, the general framework presented in Section 2 can be used to deal with several estimation problems such as dynamic model parameter estimation, prediction, filtering, etc. In this section we show how to formulate some of them, leading to polynomial S and F .

3.1. *Parameter estimation of exponential models.* Let us consider the multiexponential model

$$y(t) = \sum_{i=1}^l \mu_i e^{-\nu_i t} + e(t) \quad (18)$$

† z^* is defined as $\sup_{z \in S(T(y))} \|z^* - z\| = \inf_{z \in Z} \sup_{z \in S(T(y))} \|z - z\|$.

where μ_i and ν_i are unknown real parameters and $e(t)$ is unknown but bounded by a given $\epsilon(t)$

$$|e(t)| \leq \epsilon(t) \quad (19)$$

Suppose that m values $\{y(t_1), \dots, y(t_m)\}$ are known and the aim is to estimate parameters μ_i and ν_i , $i = 1, \dots, l$. Problems of this type arise in many applications, e.g., in pharmacokinetics and biomedical problems (Godfrey, 1983). By setting $\xi_i = e^{-\nu_i t}$, $i = 1, \dots, l$, the space Λ is the $2l$ -dimensional space of $\lambda = [\mu_1, \dots, \mu_l, \xi_1, \dots, \xi_l]'$ and $Z = \Lambda$, and S is taken as the identity operator. Y is an m -dimensional space whose elements are $\{y(t_1), \dots, y(t_m)\}' = \{y(t_1), \dots, y(t_m)\}'$.

The information operator $F(\cdot)$ is given by:

$$\begin{bmatrix} F_1(\lambda) \\ \vdots \\ F_m(\lambda) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^l \mu_i \xi_i^{t_1} \\ \vdots \\ \sum_{i=1}^l \mu_i \xi_i^{t_m} \end{bmatrix} \quad (20)$$

where it is apparent that each component of $F(\lambda)$ is a polynomial function of μ_i and ξ_i . Note that in this way variables ξ_i s are considered instead of the ν_i s. Estimates of the ν_i s and relative UI can be obtained by logarithmic transformation of the ξ_i s.

3.2. *Parameter estimation of ARMA models.* Let us consider the ARMA model

$$y_k = \sum_{i=1}^p \delta_i y_{k-i} + \sum_{i=1}^q \theta_i u_{k-i} + e_k \quad (21)$$

where e_k and u_k are unknown but bounded sequences

$$|e_k| \leq \epsilon_k, \quad |u_k| \leq \beta_k, \quad \forall k. \quad (22)$$

To keep notation as simple as possible, consider the case $p = q$. Suppose that m values $\{y_1, \dots, y_m\}$ are known and the aim is to estimate parameters δ_i, θ_i . Λ can be defined as the $2p + m - 1$ -dimensional space with elements

$$\lambda = [\delta_1, \dots, \delta_p, \theta_1, \dots, \theta_p, u_1, \dots, u_{m-1}]' \quad (23)$$

and the subset K of Λ is defined by (22). Z is the $2p$ -dimensional space with elements

$$z = [\delta_1, \dots, \delta_p, \theta_1, \dots, \theta_p]'. \quad (24)$$

The operator $S(\lambda)$ is linear and is given by

$$S(\lambda) = [I_{2p} \mid \theta] \lambda \quad (25)$$

where I_{2p} is the identity matrix of dimension $(2p, 2p)$ and θ is the null matrix of dimension $(2p, m-1)$. Y is an $m-p$ dimensional space with elements $y = \{y_{p+1}, \dots, y_m\}'$. The information operator $F(\cdot)$ is given by

$$\begin{bmatrix} F_1(\lambda) \\ \vdots \\ F_{m-p}(\lambda) \end{bmatrix} = \begin{bmatrix} \delta_1 y_p + \dots + \delta_p y_1 + \theta_1 u_p + \dots + \theta_p u_1 \\ \vdots \\ \delta_1 y_{m-p+1} + \dots + \delta_p y_{m-p} + \theta_1 u_{m-p+1} + \dots + \theta_p u_{m-p} \end{bmatrix}. \quad (26)$$

As it can be easily checked, $S(\lambda)$ is linear and $F(\lambda)$ is polynomial (actually linear in δ_i and bilinear in θ_i and u_i).

It is worth noting that the same technique can be used to deal with more general models such as ARMAX, bilinear, quadratic, etc.

3.3. *Multistep prediction with ARMA models.* Consider the ARMA model (21) and suppose that the aim is to estimate y_{m+h} when past values $\{y_1, y_m\}$ are known (*h-step ahead prediction problem*). This problem can be embedded in the framework of Section 2 by defining all spaces and functions as for the case of ARMA parameter estimation, except for Λ , Z and $S(\lambda)$. For the sake of notation simplicity, consider the case $h = 2$. The space Λ is a $2p + m + 3$ -dimensional space with elements

$$\lambda = [\delta_1, \dots, \delta_p, \theta_1, \dots, \theta_p, u_1, \dots, u_{m+1}, e_{m+1}, e_{m+2}]' \quad (27)$$

Z is the one dimensional space with elements $z = y_{m+2}$. The operator $S(\cdot)$ is no longer linear and is given by

$$S(\lambda) = \sum_{i=1}^p (\delta_i \delta_i - \delta_{i+1}) y_{m+1} + \sum_{i=1}^p (\delta_i \theta_i + \theta_{i+1}) u_{m+1} + \delta_1 e_{m+1} + e_{m+2} \tag{28}$$

where

$$\delta_i = 0, \quad \theta_i = 0, \quad \text{for } i > p.$$

The operator $S(\cdot)$ is no longer linear and is given by

$$S(\lambda) = \sum_{i=1}^p (\delta_i \delta_i - \delta_{i+1}) y_{m+1} + \sum_{i=1}^p (\delta_i \theta_i + \theta_{i+1}) e_{m+1} + e_{m+2} \tag{28}$$

where

$$\delta_i = 0, \quad \theta_i = 0, \quad \text{for } i > p \text{ and } \theta_0 = 1.$$

Note that the evaluation of the expression of $S(\lambda)$ requires symbolic computations which for large h may become cumbersome. If necessary, such symbolic computations may be performed by symbolic manipulation codes like MACSYMA, REDUCE, muSIMP, etc

3.4. *Discrete time state space models.* Let us consider the h th order linear discrete time dynamic model

$$\begin{aligned} x_{k+1} &= A(p)x_k + B(p)u_k \\ y_k &= C(p)x_k + e_k \end{aligned} \tag{29}$$

where the system matrices entries are polynomial functions of physical unknown parameters $p \in R^l$, u_k is a known sequence and e_k is an unknown but bounded sequence. Suppose, for ease of presentation, that the system is single output and that m values of the output $[y_1, \dots, y_m]$ are known. The aim is to estimate unknown parameters p and system initial condition $x_0 \in R^h$. The problem can be embedded in the framework of Section 2 as follows. The space Λ is identified as the $l+h$ -dimensional space of vectors $\lambda = [p \mid x_0]^T$; $K = \Lambda$ (if no *a priori* information is available on physical parameters and initial condition); $Z = \Lambda$ and S is identity. Y is an m -dimensional space and $F(\lambda)$ is the m -dimensional vector function given by

$$\begin{aligned} F_1(\lambda) &= C(p)A(p)x_0 + C(p)B(p)u_0 \\ &\vdots \\ F_m(\lambda) &= C(p)A^m(p)x_0 + \sum_{i=0}^{m-1} C(p)A^{m-i-1}(p)B(p)u_i \end{aligned} \tag{30}$$

The operator $F(\cdot)$ is again polynomial in the parameter vector p and linear in the initial condition x_0 . Note that symbolic computation of the polynomial expressions of $F_i(\lambda)$ in (30) is required. For large values of m , symbolic evaluation of (30) may become cumbersome due to the fast increase of the number of terms in each component of $F(\lambda)$.

4. An algorithm for the exact computation of solution uncertainty intervals

If $S(\cdot)$ and $F(\cdot)$ are polynomial functions, optimization problems (14) are of the form

$$\min (\max) f_0(\lambda) \tag{31}$$

subject to

$$|f_i(\lambda)| \leq \epsilon_i, \quad i = 1, \dots, m$$

where functions $f_i(\lambda)$ have the structure

$$f_i(\lambda) = \sum_{k=1}^n \alpha_{ik} \lambda_1^{a_{k1}} \lambda_2^{a_{k2}} \dots \lambda_n^{a_{kn}} \tag{32}$$

For example, in parameter estimation of exponential models of Section 3.1 one of the problems to be solved is

$$= \min \mu_1 \tag{33}$$

subject to

$$y(t_j) - \sum \mu_j \xi_j^t \leq \epsilon(t_j), \quad j = 1, \dots, m.$$

The above optimization problem can be transformed into a

signomial programming problem (see e.g. Ecker, 1980). Such problems are in general not convex and may exhibit local extrema. In the following we present an algorithm, due in its original version to Falk (1973), which guarantees convergence to the global extremum. The iterative algorithm allows one to evaluate upper and lower bounds on the absolute extremum at each iteration. The sequences of upper and lower bounds converge monotonically to the global solution.

A signomial optimization problem is defined as follows

$$\min (h_0(\lambda) - g_0(\lambda)) \tag{34}$$

subject to

$$\begin{aligned} h_k(\lambda) - g_k(\lambda) &\leq 1, \quad k = 1, \dots, 2m \\ \lambda_i &> 0, \quad i = 1, \dots, n \end{aligned}$$

where $h_k(\lambda)$ and $g_k(\lambda)$ ($k = 0, \dots, 2m$) are *posynomials*, i.e. polynomials with non-negative coefficients such that

$$\begin{aligned} h_k(\lambda) &= \sum_{\alpha \in I_1(k)} \alpha_i \prod_{j=1}^n \lambda_j^{a_{ij}}, \quad k = 0, \\ g_k(\lambda) &= \sum_{\alpha \in I_2(k)} \alpha_i \prod_{j=1}^n \lambda_j^{a_{ij}} \end{aligned} \tag{35}$$

where exponents a_{ij} are real numbers, α_i are positive reals and $I_{1(2)}(k)$, $k = 0, \dots, 2m$ are sets of integers disjoint for each k . Note that functions $h_k(\lambda)$ and $g_k(\lambda)$ are in general not convex. Nevertheless, we can introduce new variables x_i with the aim of transforming h_k and g_k into convex functions of the variables x_i ,

$$\lambda_i = e^{x_i}, \quad i = 1, \tag{36}$$

Equations (35) become

$$\begin{aligned} H_k(x) &= [h_k(\lambda)]_{\lambda_i = e^{x_i}} = \sum_{\alpha \in I_1(k)} \alpha_i e^{(a_i, x)} \\ G_k(x) &= [g_k(\lambda)]_{\lambda_i = e^{x_i}} = \sum_{\alpha \in I_2(k)} \alpha_i e^{(a_i, x)} \end{aligned} \tag{37}$$

where (\cdot, \cdot) denotes inner product and $a_i = [a_{i1}, \dots, a_{in}]^T$. Problem (34) is transformed into the equivalent problem

$$\min (H_0(x) - G_0(x)) \tag{38}$$

subject to

$$H_k(x) - G_k(x) \leq 1, \quad k = 1, \dots, 2m.$$

Since $H_k(x)$ (and $G_k(x)$) are convex functions, it follows that if $I_2(k) = \emptyset \forall k$, then (38) is a convex problem and the original problem (34), called a *posynomial* problem, is equivalent to a convex program.

The algorithm given below generates a tree whose nodes are associated with convex problems Q^* which approximate the signomial problem (38) (called P). Problems Q^* are obtained by suitable linear overestimates of $G_k(x)$ as follows.

Suppose that *a priori* upper and lower bounds x^m and x^M of a global solution x^* of (38) are given

$$x_j^m \leq x_j^* \leq x_j^M, \quad j = 1, \dots, n \tag{39}$$

and that

$$f^* = H_0(x^*) - G_0(x^*)$$

is the global minimum of (38). Let \mathcal{F}^* be the set defined as

$$\mathcal{F}^* = \{x: r_i^* \leq (a_i, x) \leq R_i^*, \quad i \in I_2(k)\}. \tag{40}$$

Variables r_i^* and R_i^* are recursively computed using the rules of steps 5 and 6 below, starting from the initial values

$$r_i^1 = \sum_{j=1}^n \min \{a_{ij} x_j^m, a_{ij} x_j^M\}, \quad i \in I_2(k) \tag{41}$$

$$R_i^1 = \sum_{j=1}^n \max \{a_{ij} x_j^m, a_{ij} x_j^M\}, \quad i \in I_2(k) \tag{42}$$

Approximating problems Q^* are of the form

$$\min (H_0(x) - L_0^*(x)) \tag{43}$$

subject to

$$\begin{cases} H_k(x) - L_k^*(x) \leq 1, & k = 1, \dots, 2m \\ x_j^m \leq x_j \leq x_j^M, & j = 1, \dots, n \end{cases}$$

TABLE 1. DATA FOR EXAMPLE 1

k	1	2	3	4	5	6	7	8	9	10
t_k	0.75	1.5	2.25	3.0	6.0	9.0	13.0	17.0	21.0	25.0
$y_k = y(t_k)$	7.39	4.09	1.74	0.097	-2.57	-2.71	-2.07	-1.44	-0.98	-0.66

where

$$L_k^i(x) = \sum_{i \in I_2(k)} \left(\frac{\alpha_i}{R_i^i - r_i^i} \right) [(R_i^i e^{r_i^i} - r_i^i e^{R_i^i}) + (e^{R_i^i} - e^{r_i^i})(a_i, x)] = \sum_{i \in J(s)} \mathcal{F}_i^i(x) \quad (44)$$

Note that since the terms $L_k^i(x)$ are linear and functions $H_k(x)$ are convex, problems Q^s are convex. Global solutions x^s for these problems, with minimum $v^s = H_0(x^s) - L_0(x^s)$, can be found by any convex optimization algorithm. Also notice that $L_k^i(x) \geq G_k(x) \forall x \in \mathcal{F}^s$, and consequently if $x^s \in \mathcal{F}^s$, then $v^s \leq f^*$.

The algorithm generates new approximating problems, by selecting an existing node τ , according to step 3 of the algorithm below, and refining the linear approximation of the corresponding problem Q^s , according to rules of steps 5 and 6. Only two problems are generated at each stage, so that after stage s has been completed, problems $Q^1, Q^2, \dots, Q^{2s+1}$ are generated.

Let $J(s)$ be the set of all nodes τ which have not been selected as branching nodes at stages preceding stage s (see step 3 of the algorithm below). Define V^s and U^s as

$$V^s = \min_{\tau \in J(s)} v^{\tau} \quad (45)$$

$$U^s = \min_{\tau = 1, \dots, 2s-1} \{H_0(x^{\tau}) - G_0(x^{\tau})\} \quad (46)$$

Note that approximations of functions $G_k(x)$ are performed by constructing linear envelopes, so that the minima of the two approximating problems generated at each stage s are larger than the minimum of the problem which generated them at stage $s - 1$. This guarantees that the sequence of lower bounds V^s to the global minimum never decreases. Moreover, the way in which the upper bounds U^s are generated ensures that they form a non increasing sequence. More importantly, using the results in Soland (1971), it is shown in Falk (1973) that the sequence x^s contains a subsequence converging monotonically to the global solution x^* and

$$\lim_{s \rightarrow \infty} V^s = \lim_{s \rightarrow \infty} U^s = f^* \quad (47)$$

The algorithm consists of the following steps:

Step 1. Initialization.

Generate and solve Q^1 , obtaining x^1, v^1, V^1, U^1 . Set $s = 1, \tau = 1, J(s) = \{1\}$

Step 2. Check for solution.

If $V^s = U^s$ then a global solution of problem P is

$$x^* = x^s; \quad f^* = V^s \quad (48)$$

Otherwise go to Step 3.

Step 3. Choose a branching node τ .

Select $\tau \in J(s)$ such that $v^{\tau} = V^s$

Step 4. Choose a term of $G_k(x)$ to be approximated.

Select $k^* \in \{0, 1, \dots, 2m\}$ maximizing $L_{k^*}^i(x^{\tau}) - H_{k^*}(x^{\tau})$.

Select $i^* \in I_2(k^*)$ maximizing $\mathcal{L}_{i^*}^{\tau}(x^{\tau}) - \alpha_i e^{(a_i, x^{\tau})}$.

Step 5. Generate problem Q^{2s} .

Set

$$\begin{aligned} r_i^{2s} &= r_i^{\tau}; \quad R_i^{2s} = R_i^{\tau}, \quad \forall i \in I_2(k^*), \quad i \neq i^* \\ R_{i^*}^{2s} &= (a_{i^*}, x^{\tau}), \quad r_{i^*}^{2s} = r_{i^*}^{\tau}. \end{aligned} \quad (49)$$

Step 6. Generate problem Q^{2s+1} .

Set

$$\begin{aligned} r_i^{2s+1} &= r_i^{\tau}, \quad R_i^{2s+1} = R_i^{\tau}, \quad \forall i \in I_2(k^*), \quad i \neq i^* \\ r_{i^*}^{2s+1} &= (a_{i^*}, x^{\tau}), \quad R_{i^*}^{2s+1} = R_{i^*}^{\tau}. \end{aligned} \quad (50)$$

Step 7. Solve problems Q^{2s} and Q^{2s+1}

Solve problems Q^{2s} and Q^{2s+1} , obtaining $x^{2s}, x^{2s+1}, v^{2s}, v^{2s+1}$. Compute V^{s+1}, U^{s+1} according to (45) and (46). Update the set $J(s)$: add the two nodes $\tau = 2s, \tau = 2s + 1$ and delete the node τ selected at Step 3. Set $s = s + 1$ and go to step 2.

We conclude this section by making some considerations on the estimation algorithm proposed above.

Remark 1. Computation of U^s may be improved by using in (46) a local solution to the true problem P , computed by an iterative algorithm starting from x^1 , instead of using $\{H_0(x^{\tau}) - G_0(x^{\tau})\}$.

Remark 2. The condition $\lambda_i > 0$ in (34) is not a serious restriction. In fact, it is possible to bring the set $T(y)$ in the first orthant of Λ by means of a suitable translation of the origin of the problem element space. Another way of dealing with this problem is to express unknown sign variables as differences of auxiliary strictly positive variables.

Remark 3. The convergence speed of the algorithm is in general quite sensitive to the sizes of intervals $x_i^M - x_i^m$. In solving any of the 2I optimization problems (14), information gained by the solved ones can be used to shrink as much as possible such intervals. This is particularly simple for parameter estimation problems where $z_i = (S(\lambda))_i = \lambda_i$. In fact, the following heuristic strategy can be used for handling this problem. A certain number of runs of the 2I optimization problems (14) are performed, stopping the algorithm after few stages \bar{s} (say $\bar{s} = 5$), without waiting for convergence of upper and lower bounds. When solving the first problem of (14), i.e. finding $z_1^m = \mu_1^m, x_1^m$ and x_1^M can be derived by *a priori* information provided by the set K . When solving the second problem, i.e. computation of z_1^M in (14), we can set (recall (36)) $x_1^m = \ln V_1$, where V_1 is the lower bound of z_1^m obtained by the preceding run of the algorithm stopped at stage \bar{s} . In solving the third problem (computation of z_2^m), we can set $x_1^M = \ln U_1$, U_1 is the upper bound of z_1^M provided by the preceding run of the algorithm stopped at stage \bar{s} , and so on. This procedure is iterated until it is able to tighten the bounds x_i^m or x_i^M . Successively, the limitation on the number of stages is removed and each extremum problem is solved by letting the algorithm reach convergence.

Such a shrinking procedure has been used in working out the numerical examples reported in the next section, and proved to be very effective, leading to considerable computing time reductions

5. Numerical examples

Example 1. Parameter estimation of a multieponential model.

The following model is considered

$$y(t) = \mu_1 e^{-\nu_1 t} + \mu_2 e^{-\nu_2 t} + e(t). \quad (51)$$

The data used are reported in Table 1. They have been generated from (51) with the following nominal parameter values

$$\mu_1 = 20.0, \quad \nu_1 = 0.4, \quad \mu_2 = -8.0, \quad \nu_2 = 0.1. \quad (52)$$

TABLE 2. UNCERTAINTY INTERVALS AND CENTRAL ESTIMATES FOR EXAMPLE 1

	μ_1	ν_1	μ_2	ν_2
UIs	[17.2, 26.9]	[0.30, 0.49]	[-16.1, -5.4]	[0.077, 0.136]
Central estimates	22.05	0.395	-10.75	0.1065

TABLE 3. DATA FOR EXAMPLE 2

k	1	2	3	4	5	6	7	8	9	10	11	12
y_k	0.19	-0.72	-0.82	-0.22	0.88	0.80	-0.20	-0.88	0.31	0.32	-0.33	-0.63

TABLE 4. UNCERTAINTY INTERVALS AND CENTRAL PREDICTIONS FOR EXAMPLE 2

	y_{13}	y_{14}	y_{15}	y_{16}
UI _k	[-0.53, 0.56]	[-0.33, 1.21]	[-0.875, 1.19]	[-1.75, 0.87]
Central prediction	0.01	0.44	0.16	-0.44
Nominal prediction	0.04	0.45	0.11	-0.28

The bounds on measurement errors are supposed to be:

$|e(t_k)| \leq 0.05 |y(t_k)| + 0.1.$ (53)

A priori information set K is defined by the following inequalities

$$K: \begin{cases} 2.0 \leq \mu_1 \leq 60.0 \\ 0.0 < v_1 \leq 1.0 \\ -30.0 \leq \mu_2 \leq -1.0 \\ 0.0 < v_2 \leq 0.5. \end{cases}$$
 (54)

The estimation results obtained are reported in Table 2. They refer to convergence within 2% of upper and lower bounds of the signomial algorithm for each extremization problem (14).

The total computing time of the algorithm, using the shrinking procedure outlined in Remark 3 of Section 4, is about 10 minutes on a VAX 8800 computer. Convergence within the mentioned tolerance, without using the shrinking procedure, has not been reached after a computing time of about one order of magnitude larger.

Example 2. Multistep prediction with an AR model.

The following AR(2) model is considered

$y_k = \delta_1 y_{k-1} + \delta_2 y_{k-2} + e_k$ (55)

The data used, which are reported in Table 3, have been generated from (55) with the nominal parameter values

$\delta_1 = 0.3, \quad \delta_2 = -0.69$ (56)

assuming e_k uniformly distributed and such that

$|e_k| \leq 0.5.$ (57)

Multistep predictions from 1 to 4 steps ahead have been computed by considering as a priori information the set K defined by the following inequalities

$$K: \begin{cases} 0.19 \leq \delta_1 \leq 0.4 \\ -0.8 \leq \delta_2 \leq -0.51. \end{cases}$$
 (58)

The uncertainty intervals in (58) have been obtained by a preliminary analysis of maximal and minimal feasible parameters of the linear model (55) by means of linear programming.

The results obtained are reported in Table 4. They refer to convergence within 2% of upper and lower bounds of the signomial algorithm for each extremization problem (14). The last line of Table 4 reports the predictions (called nominal predictions) obtained by the minimum mean square predictor of (55) with the nominal parameter values (56).

The total computing time for obtaining these results is approximately 3 minutes on a VAX 8800 computer.

6. Conclusions

A method has been proposed for parameter estimation and prediction in a set membership uncertainty context, when dynamic models are nonlinear in the variables to be estimated. A procedure has been presented which allows one to compute exact uncertainty intervals of the estimated variables for the case when measurements depend polynomially on model parameters. Some examples have been worked out to show applications of the proposed algorithm.

Acknowledgements—This work was partially supported by funds of Ministero della Università e della Ricerca Scientifica e Tecnologica and Camera di Commercio di Torino.

References

Belforte, G. and B. Bona (1985). An improved parameter identification algorithm for signals with unknown but bounded errors. *Proc. 7th IFAC Symp. on Identification and System Parameter Estimation*, York, pp. 1507-1511.

Belforte, G. and M. Milanese (1981). Uncertainty intervals evaluation in presence of unknown but bounded errors. Nonlinear families of models. *Proc. 1st IASTED Symp. on Modelling, Identification and Control*, Davos, pp. 75-79.

Clement T. and S. Gentil (1988). Reformulation of parameter identification with unknown but bounded errors. *Math. Comput. Simulat.* **30**, 257-270.

Ecker, J. G. (1980). Geometric programming: Methods, computations and applications. *SIAM Rev.*, **1**, 339-362.

Falk, J. E. (1973). Global solutions of signomial programs. Tech. Rep. T-274, George Washington Univ., Washington, DC.

Fogel, E. and F. Huang (1982). On the value of information in system identification—bounded noise case. *Automatica*, **18**, 140-142.

Godfrey, K. (1983). *Compartmental Models and Their Applications*. Academic Press, NY.

Kaciewicz, B. Z., M. Milanese, R. Tempo and A. Vicino (1986). Optimality of central and projection algorithms for bounded uncertainty. *Syst. Control Lett.*, **8**, 161-171.

van Laarhoven, P. J. M. and E. H. Aarts (1987). *Simulated Annealing: Theory and Applications*. Reidel, Dordrecht.

Micchelli, C. A. and T. J. Rivlin (1977). A survey of optimal recovery. In Micchelli, C. A. and T. J. Rivlin (Eds), *Optimal Estimation in Approximation Theory*. Plenum, New York, pp. 1-54.

Milanese, M. (1989). Estimation theory and prediction in the presence of unknown but bounded uncertainty: A survey. In Milanese, M., R. Tempo and A. Vicino (Eds), *Robustness in Identification and Control*. Plenum Press, NY.

Milanese, M. and G. Belforte (1982). Estimation theory and uncertainty intervals evaluation in presence of unknown but bounded errors. Linear families of models and estimators. *IEEE Trans. Aut. Control*, **AC-27**, 408-414.

Milanese, M. and R. Tempo (1985). Optimal algorithms theory for robust estimation and prediction. *IEEE Trans. Aut. Control*, **AC-30**, 730-738.

Milanese, M., R. Tempo and A. Vicino (1986). Strongly optimal algorithms and optimal information in estimation problems. *J. Complexity*, **2**, 78-94.

Norton, J. P. (1987). Identification and application of bounded-parameter models. *Automatica*, **23**, 497-507.

Pardalos P. M. and J. B. Rosen (1987). *Constrained Global Optimization*. Springer, Berlin.

Schweppe, F. C. (1973). *Uncertain Dynamic Systems*. Prentice-Hall, Englewood Cliffs, NJ.

Smit, M. K. (1983). A novel approach to the solution of indirect measurement problems with minimal error propagation. *Measurement*, **1**, 181-190.

Soland, R. M. (1971). An algorithm for separable nonconvex programming problems II: Nonconvex constraints. *Management Sci.*, **17**, 759-773.

Traub, J. F. and H. Woźniakowski (1980). *A General Theory of Optimal Algorithms*. Academic Press, New York.

Walter, E. and H. Piet-Lahanier (1986). Robust nonlinear parameter estimation in the bounded noise case. *Proc. 25th Conf. on Decision and Control*, Athens.

Linear Time-invariant Distributed Parameter System Identification via Orthogonal Functions*

B. M. MOHAN† and K. B. DATTA‡‡

Key Words—Parameter estimation; initial and boundary conditions estimation; distributed parameter systems; orthogonal functions

Abstract—This paper points out the mathematical inconsistencies traced in the literature on the identification problem of linear time-invariant distributed parameter systems via orthogonal functions, and proposes for the same problem a unified identification approach based on the concept of one shot operational matrix for repeated integration. It presents identifiability requirements for the block-pulse functions approach while suggesting a linear independence test for the full column rank of linear algebraic system arising out of the system model upon the application of orthogonal functions. Finally, it illustrates system identification with a numerical example.

1. Introduction

THE AIM of this paper is to develop a general and unified identification approach using orthogonal functions (OF) for the estimation of parameters, initial conditions (ICs) and boundary conditions (BCs) of linear time-invariant single-input single-output continuous-time distributed parameter systems (DPS) modelled by

$$a_n \frac{\partial^2 y(x, t)}{\partial t^2} + a_{n1} \frac{\partial^2 y(x, t)}{\partial x \partial t} + a_{n2} \frac{\partial^2 y(x, t)}{\partial x^2} + a_n \frac{\partial y(x, t)}{\partial t} + a_{n1} \frac{\partial y(x, t)}{\partial x} + a_n y(x, t) = u(x, t) \quad (1)$$

from its input-output records available over the region $x \in [x_0, x_f]$, $t \in [t_0, t_f]$. Although some attempts have already been made on this problem in the past, they seem to have the following mathematical inconsistencies.

As it appears, Paraskevopoulos and Bounas (1978) were the first authors to investigate this problem via Walsh functions (WF). To estimate the parameters with the ICs,

$$f(x) = y(x, t_0), \quad g(x) = \left. \frac{\partial y(x, t)}{\partial t} \right|_{t=t_0}$$

and the BCs:

$$q(t) = y(x_0, t), \quad r(t) = \left. \frac{\partial y(x, t)}{\partial x} \right|_{x=x_0}$$

they integrated (1) twice with respect to x and twice with respect to t to obtain an integral equation, approximated all the functions in x and/or t in finite Walsh series, and arrived at the following matrix algebraic equation from the integral equation by employing the integral operational property of WF.

$$\begin{aligned} & a_n (E_1^1)^2 Y + a_{n1} Y E_1^1 + a_{n1} E_1^1 Y E_1^1 + a_n (E_1^1)^2 Y E_1^1 \\ & + a_n E_1^1 Y E_1^1 + a_n (E_1^1)^2 Y E_1^1 = a_n (E_1^1)^2 \sum_{j=0}^{n-1} f_j \Delta_{j+1,1} \\ & + a_{n1} \sum_{j=0}^{n-1} q_j \Delta_{j+1,1} E_1^1 + (E_1^1)^2 \sum_{j=0}^{n-1} h_j \Delta_{j+1,1} E_1^1 \\ & + a_{n1} (E_1^1)^2 \sum_{j=0}^{n-1} w_j \Delta_{j+1,1} E_1^1 + E_1^1 \sum_{j=0}^{n-1} z_j \Delta_{j+1,1} E_1^1 \\ & = (E_1^1)^2 U E_1^1 \end{aligned}$$

with $h(x) = a_n g(x) + a_{n1} f(x)$, $w(x) = \partial y(x, t_0) / \partial x$ and $r(t) = a_{n1} q(t) + a_{n1} r(t) + a_{n1} \int_{t_0}^t q(\tau) d\tau$. Then they rewrote the above algebraic equation in the form of $M\mathbf{p} = \mathbf{v}$ and attempted to estimate the augmented parameter vector \mathbf{p} from the least-square technique i.e. $\hat{\mathbf{p}} = [M^T M]^{-1} [M^T \mathbf{v}]$. Here it is important to note that the columns of M corresponding to h_j and $a_{n1} w_j$ terms (shown with a broken underline in the matrix algebraic equation) are apparently linearly dependent. Due to this linear dependence of columns of M , an inverse of $[M^T M]$ does not exist making the identification impossible. Hence, it may be concluded that the algorithm described above is not always suitable for the identification of (1). Jha and Zaman (1985) used the same algorithm for identification using Laguerre polynomials (LaP). Recently, Horng *et al.* (1986) investigated the identification problem of first-order DPS via shifted Chebyshev polynomials of the first kind (CP1). Most recently, Mohan and Datta (1988, 1989) have explored the potentialities of shifted Legendre polynomials (LeP) and sine-cosine functions (SCF) in DPS identification.

Having seen the state of affairs, it appears to the authors that the problem of general second-order DPS identification has not yet been studied via block-pulse functions (BPF), CP1 and Chebyshev polynomials of the second kind (CP2). The concept of the one-shot operational matrix for repeated integration (OSOMRI) (Unbehauen and Rao, 1987) has recently been extended to shifted LeP and SCF and used, owing to its superiority, for the identification of DPS, by Mohan and Datta (1988, 1989). OSOMRIs of CP1 and CP2 are not yet reported. A comparative study of all OF approaches to assess the relative merits and demerits of each approach in DPS identification is also not yet reported in the literature.

In this paper, by taking all the aforementioned points into consideration, an investigation is therefore made on the identification problem of DPS to eliminate the lacunae of existing methods and to introduce a more general unified identification approach based on the concept of OSOMRI. This paper, presenting some important results of this investigation, is organized as follows. In Section 2 some mathematical preliminaries including OSOMRI of CP1 and CP2 are presented. In Section 3 we present a unified identification approach for any class (elliptic, parabolic or hyperbolic) of DPS. Section 4 discusses the identifiability requirements, which are followed by a demonstration of the proposed identification scheme and a comparative study of all OF approaches by an illustrative example.

* Received 3 May 1989; 15 January 1990; received in final form 11 July 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. Bitmead under the direction of Editor P. C. Parks.

† Department of Electrical Engineering, Indian Institute of Technology, Kharagpur 721 302, India.

‡ Author to whom all correspondence should be addressed.

2. Mathematical preliminaries

All the finite range orthogonal polynomials such as CP1, CP2 and LeP are orthogonal over $z \in [-1, 1]$. The shifted polynomials, orthogonal over $t \in [t_0, t_f]$, may be obtained by setting

$$z = 2(t - t_0)/(t_f - t_0) - 1. \tag{2}$$

It is well known that a two-variable function $f(x, t)$ square-integrable in the region $x \in [x_0, x_f], t \in [t_0, t_f]$ can be approximately expanded in a series of OF as

$$f(x, t) = \sum_{i=0}^m \psi_i(x) \sum_{j=0}^n f_{ij} \phi_j(t) = \Psi^T(x) F \Phi(t) \tag{3}$$

where $\Psi(x)$ is an m -dimensional OF(basis) vector in x , $\Phi(t)$ is an n -dimensional basis vector in t and F is an $m \times n$ matrix, commonly known as a spectrum of $f(x, t)$, whose elements f_{ij} are given by

$$f_{ij} = c_1 c_2 \int_{t_0}^{t_f} \int_{x_0}^{x_f} w_1(x) w_2(t) f(x, t) \psi_i(x) \phi_j(t) dx dt \tag{4}$$

in which $w_1(x)$ and $w_2(t)$ are the weight functions of orthogonal system and c_1 and c_2 are the constants. Weight functions $w(\cdot)$ and constants c are different for different orthogonal systems [see for instance Unbehauen and Rao (1987) for BPF and CP2; Horng *et al.* (1986) for CP1; Mohan and Datta (1988 and 1989) for LeP and SCF; and Jha and Zaman (1985) for LaP].

Since for CP1, $w_1(x)$ and $w_2(t)$ are singular at the end points, one may have to use open quadrature formulae, e.g. the Gauss-Chebyshev first kind open quadrature formula for the computation of f_{ij} . This requires explicit knowledge of $f(x, t)$. In the actual situation, $f(x, t)$ is known either continuously or at some discrete points (preferably equispaced) in the region of measurements. When the signal $f(x, t)$ is purely deterministic and is available only at some points, computation of f_{ij} is still possible by interpolating $f(x, t)$ at the zeros of CP1 used in the numerical integration. However, when $f(x, t)$ is corrupted with noise computation of f_{ij} is not possible as interpolating noisy signal is meaningless.

Using the concept of OSOMRI we have

$$\int_{t_0}^{t_f} \int_{x_0}^{x_f} \int_{t_0}^{t_f} \int_{x_0}^{x_f} f(x, t) dx^2 dt^2 = \Psi^T(x) E_{12}^T F E_{12} \Phi(t) \tag{5}$$

where E_{12} and E_{12}^T are $m \times m$ and $n \times n$ OSOMRIs of $\Psi(x)$ and $\Phi(t)$ respectively. $E_{11} = E_x$ and $E_{11}^T = E_x^T$ are the well known integration operational matrices of the same dimension (Unbehauen and Rao, 1987). The derivation of OSOMRIs of BPF, LeP and SCF is given in Unbehauen and Rao (1987) and Mohan and Datta (1988 and 1989). For LaP, OSOMRI is the same as the conventional repeated integration operational matrix.

OSOMRI of shifted CP1 and CP2. Here we present the expressions for generating the elements of OSOMRI of CP2. Shifted CP2 have the following recurrence relations:

$$2(i + 2)\phi_i(t) = [-\phi_i(t)\phi_i^{(1)}(t)/2 + \phi_i^{(1)}(t)](t_f - t_i) \tag{6}$$

$$\phi_i(t)\phi_i(t) = \phi_{i-1}(t) + \phi_{i+1}(t), \quad i \geq 1 \tag{7}$$

and

$$\phi_i(t_0) = (-1)^i(i + 1). \tag{8}$$

Integrating (6) twice with respect to t gives

$$\begin{aligned} \int_{t_0}^{t_f} \int_{t_0}^{t_f} \phi_i(\tau) d\tau^2 &= \frac{(t_f - t_0)^2}{4} \left\{ \frac{(-1)^i(i + 1)}{i(i + 2)} \phi_0(t) + \frac{(-1)^i}{2(i + 1)} \phi_1(t) \right. \\ &\quad \left. + \frac{\phi_{i-2}(t)}{4i(i + 1)} - \frac{\phi_i(t)}{2i(i + 2)} + \frac{\phi_{i+2}(t)}{4(i + 1)(i + 2)} \right\} \text{ for } i = 2, 3, 4, \dots \end{aligned}$$

where (7) and (8) are used for simplification. For $i = 0$ and 1 we have

$$\begin{aligned} \int_{t_0}^{t_f} \int_{t_0}^{t_f} \phi_0(\tau) d\tau^2 &= (t_f - t_0)^2 [0.625\phi_0(t) + 0.5\phi_1(t) + 0.125\phi_2(t)]/4 \end{aligned}$$

and

$$\begin{aligned} \int_{t_0}^{t_f} \int_{t_0}^{t_f} \phi_1(\tau) d\tau^2 &= (t_f - t_0)^2 [-2\phi_0(t)/3 - 5\phi_1(t)/12 + \phi_2(t)/24]/4. \end{aligned}$$

Similarly, following in the same lines, it is quite possible to arrive at the similar set of expressions for CP1 also.

3. Unified identification approach

Depending upon the values of parameters a_n, a_{11}, a_{12} , the system (1) may become elliptic, parabolic or hyperbolic. We first obtain an integral equation by integrating equation (1) twice with respect to x and twice with respect to t ; approximate all known and unknown functions in x and/or t in a finite set of OF and introduce the same in integral equation; make use of equation (5); and simplify to obtain

$$Q\theta = v \tag{9}$$

where

$$\begin{aligned} Q = & [\text{vec}(E_{12}^T Y) \mid \text{vec}(YE_{12}) \mid \text{vec}(E_{11}^T YE_{11}) \mid \text{vec}(E_{12}^T YE_{12}) \\ & \mid \text{vec}(E_{11}^T YE_{12}) \mid \text{vec}(E_{12}^T YE_{12}) \mid \dots \mid \text{vec}(E_{12}^T \Delta_{11}) \mid \dots \\ & \mid \text{vec}(E_{12}^T \Delta_{\alpha 1}) \mid \text{vec}(\Delta_{11} E_{12}) \mid \dots \mid \text{vec}(\Delta_{1n} E_{12}) \\ & \mid \text{vec}(E_{11}^T \Delta_{11} E_{11}) \mid \text{vec}(E_{12}^T \Delta_{11} E_{11}) \mid \dots \mid \\ & \text{vec}(E_{12}^T \Delta_{\gamma 1} E_{11}) \\ & \mid \text{vec}(E_{11}^T \Delta_{11} E_{12}) \mid \dots \mid \text{vec}(E_{12}^T \Delta_{1n} E_{12})] \dots \end{aligned} \tag{10}$$

$$\theta = [a_n, a_{11}, a_{12}, a_t, a_x, a_{11}, a_{12}, a_{13}, \dots, \hat{f}_n, \dots, \hat{f}_{n-1}, \hat{q}_0, \dots, \hat{q}_{n-1}, \hat{c}, h_0, \dots, h_{n-1}, \delta_0, \dots, \delta_{n-1}]^T \tag{11}$$

$$v = \text{vec}(E_{12}^T U E_{12}) \tag{12}$$

$$\hat{c} = a_{11} y(x_0, t_0), \hat{f}_i = a_n f_i, \hat{q}_i = a_{11} q_i \tag{13}$$

$$h(x) = a_n g(x) + a_{11} df(x)/dx + a_{12} f(x) \tag{14}$$

$$s(t) = a_{11} r(t) + a_{12} dq(t)/dt + a_{13} q(t) \tag{15}$$

$\alpha \leq m, \beta \leq n, \gamma \leq m, \delta \leq n, \Delta_{ij}$ is an $m \times n$ matrix having (i, j) th element unity and all other elements zero, and $\text{vec}(\cdot)$ is the vector valued function of matrix (\cdot) . Since there are mn equations and $(7 + \alpha + \beta + \gamma + \delta)$ unknowns in equation (9), the rank of Q must be equal to the number of unknowns to determine θ from

$$\hat{\theta} = [Q^T Q]^{-1} [Q^T v]. \tag{16}$$

Otherwise identification is not possible. Once θ is obtained, the initial and boundary conditions $f(x)$ and $q(t)$ can also be obtained from the equation (13). For finding the other IC $g(x)$, we obtain an integral equation by integrating equation (14) once with respect to x , approximate all the functions in x in finite set of OF and introduce the same in the integral equation, make use of integration operational property of OF and simplify. This results in

$$g(x) = \Psi^T(x) \{ [h - a_{11} f] + a_{11} (E_{11}^T)^{-1} [c e_1 - f] \} / a_n \tag{17}$$

with

$$e_1 = [1, 0, \dots, 0]^T \text{ an } m\text{-vector.} \tag{18}$$

In the same manner the other BC $r(t)$ in equation (15) may also be obtained from

$$r(t) = \{ [s^1 - a_{12} q^1] + a_{12} [c e_1^1 - q^1] E_{11}^{-1} \Phi(t) / a_{11} \} \tag{19}$$

with

$$e_1 = [1, 0, \dots, 0]^T \text{ an } n\text{-vector} \tag{20}$$

Thus the present algorithm is capable of estimating parameters as well as ICs and BCs of the system (1). This approach may be applied via WF and all classes of orthogonal polynomials. Since WF have to be chosen on the basis of 2^K where K is a positive integer, identification via WF normally becomes computationally laborious. In order to apply this algorithm via BPF and SCF certain terms in it must be redefined in the manner as discussed in the following two subsections.

TABLE 1. CONDITIONS UNDER WHICH BPF APPROACH WORKS

Condition	c	$f(x)$	$g(x)$	$h(x)$	$q(t)$	$r(t)$	$s(t)$
1	p	a	a	a	a	a	a
2	a	p	a	p	a	a	a
3	a	a	p	p	a	a	a
4	a	a	a	a	p	a	p
5	a	a	a	a	a	p	p

a = absent; p = present.

3.1. *Identification via BPF.* In this approach $\alpha = \gamma = m$ and $\beta = \delta = n$ always. For $i = 1, 2, \dots, m$ matrix Δ_{1i} must be replaced by Δ_{1i} which is an $m \times n$ matrix containing all i th row elements unity and all other elements zero. Similarly, for $j = 0, 1, \dots, n$, matrix Δ_{1j} must be replaced by Δ_{1j} which is also an $m \times n$ matrix but having all j th column elements unity and all other elements zero. In the column corresponding to \hat{c} term in equation (11) matrix Δ_{11} must be replaced by Δ_{11} which is again an $m \times n$ matrix but containing all elements unity. Vectors e_i and e_j in equations (18) and (20) must be considered as $e_i = [1, 1, \dots, 1]^T$ and $e_j = [1, 1, \dots, 1]^T$ with the same dimensions.

To determine θ from equation (16), the rank of Q must be equal to $7 + 2(m + n)$. This clearly shows that the size of θ is very much dependent on m and n which need not be so always in other OF approaches as the size of θ is decided by α, β, γ and δ . This point may be considered as the drawback of BPF approach.

3.2. *Identification via SCF.* It is well known that the Fourier series expansion of any square-integrable function contains $2n + 1$ terms (n cosine terms + $(n + 1)$ sine terms) for a chosen n . Hence, the dimensions of all vectors and matrices in Section 3 must be accordingly taken care of. Owing to this over dimension, the rank of Q in this approach must be equal to $3 + 2(\alpha + \beta + \gamma + \delta)$ assuming that $\alpha = \beta = \gamma = \delta \neq 0$. Moreover, this over dimension also makes identification computationally laborious compared to other OF approaches.

4. Practical limitations of unified identification approach

It was stated in the preceding section that Q must have its full column rank. This condition sometimes fails as the columns of Q turn out to be linearly dependent under certain circumstances. To investigate this important aspect we rewrite equation (9) as

$$[A \mid B] \tag{21}$$

where A contains all the columns corresponding to system parameters and B contains all the remaining columns of Q .

Now the rank of Q is always less than its number of columns if there exist at least two linearly dependent columns of A or B . Since the linearly dependent columns of A may only occur due to Y , it must be tested essentially with respect to the output signal $y(x, t)$. No general conclusions can be made in this context. To study the linearly dependent columns of B , the complete structure of B must be known first. This depends on operational matrices E_x, E_r, E_{x_2} and E_{r_2} and the values of α, β, γ and δ . Since B is totally independent of input and output signals, for the chosen

orthogonal system and α, β, γ and δ values, B must be tested separately before proceeding to the identification. Extensive computational experimentation has revealed that there are many possible combinations of $\alpha \leq m, \beta \leq n, \gamma \leq m$ and $\delta \leq n$ for which B and thereby Q turn out to have less than their full column ranks, making the identification a failure. Therefore, identification via WF, any class of orthogonal polynomials or SCF must be preceded by a linear independence test on B .

Owing to the fact that $\alpha = \gamma \leq m$ and $\beta = \delta \leq n$, it is quite possible to say something about the linearly dependent columns of B in BPF approach. Conducting the test on B , out of all possible combinations of α, β, γ and δ , five combinations only turn out to be favorable for identification (see Table 1).

Condition 1 is quite obvious as B contains only one column. Conditions 2-5 can be proved in simple mathematical terms. Operational matrices, E_x, E_r, E_{x_2} and E_{r_2} are always nonsingular. Consequently, the columns corresponding to $f(x), h(x), q(t)$ or $s(t)$ in equation (11) are always linearly independent, proving the conditions 3 and 5. Conditions 2 and 4 may be proved by formulating the columns of B and using the Gauss reduction method with column operations in each case.

It is well known that Haar functions and WF can be expressed as linear combinations of BPF. Therefore, identification via Haar functions or WF is not possible when the following two conditions are simultaneously met: (i) The number of basis functions is exactly the same as that of BPF while meeting the requirement of 2^k where k is some positive constant, and (ii) Conditions shown in Table 1 are not met.

5. Illustrative example

Suppose it is required to estimate the parameters as well as IC and BC of the system (hyperbolic) modelled by equation (1) with $a_{11} = a_{22} = 0$ when the input $u(x, t) = (xt)^2 + (xt + 1)(x + t)$ and the output $y(x, t)$ are available over the region $x, t \in [0, 16]$.

By employing the unified identification approach with $m = n = 3$ and $\gamma = \delta = 1$, all the parameters and IC and BC are estimated via all the classes of OF. The estimated results are as shown in Table 2. It may be observed that the estimates are very much improved when the value of m and n is increased from 3 to 12 in SCF approach, see the second set.

6. Conclusions

A unified approach for the identification of DPS via any class of OF is presented with its practical limitations. Disregarding the computational effort, all the classes of orthogonal polynomials seem to be powerful in DPS identification. Owing to the larger dimension and poorer convergence of Fourier series of nonsinusoidal signals, SCF approach is seen to be computationally laborious. On the whole, the computational requirements in DPS identification via any class of OF are highly demanding.

References

- Horng, I. R., J. H. Chou and C. H. Tsai (1986). Analysis and identification of linear distributed systems via Chebyshev series. *Int. J. Syst. Sci.*, **17**, 1089.
- Jha, A. N. and S. Zaman (1985). Identification of linear distributed systems using Laguerre operational matrices. *Int. J. Syst. Sci.*, **16**, 761.

TABLE 2. ESTIMATES OF PARAMETERS IC AND BC

Approach	a_{11}	a_1	a_2	a	$y(x, 0)$	$y(0, t)$
Actual	-0.5	0.5	0.5	1	1	1
BPF	failed due to the presence of IC and BC					
CP1	-0.500003	0.4999987	0.4999987	1.0000079	1	1
CP2	-0.5000039	0.499999	0.499999	1.0000089	1	1
LeP	-0.499999	0.5000002	0.5000002	0.9999978	1.0000002	1.0000002
LaP	-0.4994676	0.4993234	0.4993234	1.000663	0.9982402	0.9982402
SCF	-0.500689	0.4980849	0.5980849	1.0046791	1.135037	1.135037
	-0.5004474	0.4999774	0.4999774	1.0006251	1.0279839	1.0279839

- Mohan, B. M. and K. B. Datta (1988). Lumped and distributed parameter system identification via shifted Legendre polynomials. *Trans. ASME J. Dynam. Syst., Meas. Control*, **110**, 436.
- Mohan, B. M. and K. B. Datta (1989). Identification via Fourier series for a class of lumped and distributed parameter systems. *IEEE Trans. Cir. Syst.*, **CAS-36**, 1454.
- Paraskevopoulos, P. N. and A. C. Bounas (1978). Distributed parameter system identification via Walsh functions. *Int. J. Syst. Sci.*, **9**, 75.
- Unbehauen, H. and G. P. Rao (1987). *Identification of Continuous Systems*. Systems and Control series, 10, North-Holland, Amsterdam.

Brief Paper

Identification and Rational L^2 Approximation: A Gradient Algorithm*

LAURENT BARATCHART,^{††} MICHEL CARDELLI[†]
and MARTINE OLIVI[†]

Key Words—Approximation theory; computational methods; model reduction; identification; linear systems; system theory.

Abstract—This paper deals with the identification of linear constant dynamical systems when formalized as a rational approximation problem. The criterion is the L^2 norm of the transfer function, which is of interest in a stochastic context. The problem can be expressed as nonlinear optimization in a Hilbert space, but standard algorithms are usually not well adapted. We present a generic recursive procedure to find a local optimum of the criterion in the case of scalar systems. Our methods are borrowed from differential theory mixed with a bit of classical complex analysis. To our knowledge, the algorithm described in this paper is the first that ensures convergence to a local minimum.

1. Introduction

IN THIS paper, we approach the problem of identification within the framework of Hardy spaces by considering this as a rational approximation problem. We restrict ourselves to linear constant strictly causal single-input single-output dynamical systems (in short, systems). We first consider the case of a discrete time system. Let $f_1, f_2, \dots, f_m, \dots$ be its impulse response. Identifying the system usually means recognizing the sequence (f_m) as the Taylor coefficients at infinity of a proper rational function whose denominator degree (in irreducible form) is then the order of the system. But since such a sequence might not exist, in practice one has to be content with finding a rational sequence (r_m) that resembles (f_m) . This, of course, has no definite meaning, and some criteria has to be chosen. Such criteria can occur in connection with stability. Assume, for instance, that the system is l^k -stable for some $k \geq 1$, that is

$$\sum_{m=1}^{\infty} |f_m|^k < \infty.$$

One can then look for some (r_m) which is close to (f_m) in the l^k sense. Since increasing the order of (r_m) arbitrarily is unacceptable, it is also reasonable to bound it from above by some number n . Note that the identified model (r_m) will then automatically be stable.

Our assumptions have as an effect that the transfer function

$$f(z) = \sum_{m=1}^{\infty} f_m z^{-m},$$

as well as the rational function

$$r(z) = \sum_{m=1}^n r_m z^{-m}$$

are holomorphic for $|z| > 1$, and one can ask in which sense they are close to each other from an analytic viewpoint. In general, there is no completely satisfactory answer. A partial result in this direction involves the so-called real Hardy spaces H_μ , where $1 \leq \mu \leq \infty$, that we now define. If h is analytic for $|z| > 1$ and $\rho > 1$ is a real number, define a function h_ρ on the unit circle T by putting $h_\rho(e^{i\theta}) = h(\rho e^{i\theta})$. By definition, H_μ will consist of those h vanishing at infinity, assuming real values for real arguments and such that

$$\sup_{\rho > 1} \|h_\rho\|_\mu$$

where $\|\cdot\|_\mu$ is the norm in $L^\mu(T)$. It is then standard to show (Fuhrmann, 1981), that h has a radial limit h^* almost everywhere on T which lies in $L^\mu(T)$, whose Fourier coefficients are real and those of non-negative rank do vanish. Moreover, we have $\|h^*\|_\mu = \sup_{\rho > 1} \|h_\rho\|_\mu$. Conversely, any member of $L^\mu(T)$ with Fourier coefficients as above is the radial limit of some unique element of H_μ . One can then identify h and h^* and consider $\|\cdot\|_\mu$ as a norm on H_μ , that defines it as a Banach space.

Now, if k' is the conjugate of k ($1/k + 1/k' = 1$), the Hausdorff–Young theorem (Duren, 1970) gives

$$\text{when } 1 \leq k \leq 2, \quad \|f\|_{k'} \leq \left(\sum_{m=1}^n |f_m|^k \right)^{1/k}, \quad (1)$$

so that f belongs to $H_{k'}$ whenever (f_m) belongs to l^k , and that to be close in the l^k sense for impulse responses implies to be close in the $H_{k'}$ sense for transfer functions; and

$$\text{when } 2 \leq k \leq \infty, \quad \left(\sum_{m=1}^n |f_m|^k \right)^{1/k} \approx \|f\|_k, \quad (2)$$

so that to be close in the H_k sense for transfer functions implies to be close in the l_k sense for impulse responses.

Of course, there may be other reasons for being interested in rational approximation, even if $2 < k'$. The H_∞ norm, for instance, is the operator norm $l^2 \rightarrow l^2$. Nevertheless, rational approximation in H_k is demonstrably relevant to identification of impulse responses only if $1 \leq k' \leq 2$.

Clearly from the above, the case where $k = k' = 2$ is particularly nice. The conjunction of (1) and (2) is just Parseval's equality:

$$\|f\|_2^2 = \left(\sum_{k=1}^{\infty} |f_k|^2 \right)$$

and H_2 is a Hilbert space with scalar product

$$(f, h) = \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta}) \overline{h(e^{i\theta})} d\theta$$

* Received 7 February 1989; revised 5 February 1990; received in final form 5 May 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. Bitmead under the direction of Editor P. C. Parks.

[†] Institut National de Recherche en Informatique et Automatique, route des Lucioles, Sophia-Antipolis 06565, Valbonne Cedex, France.

^{††} Author to whom all correspondence should be addressed.

which in turn can be converted into a line integral

$$(f, h) = \frac{1}{2i\pi} \int_{\Gamma} f(z) h\left(\frac{1}{z}\right) \frac{dz}{z}.$$

In the rest of this paper, we shall restrict ourselves to rational approximation in H_2 which we phrase as follows:

For $f \in H_2$, minimize $\|f - r\|_2$ where r ranges over all rational functions in H_2 of order at most n .

Note that a rational fraction $r = p/q$ belongs to H_2 if and only if $\deg(p) < \deg(q)$ and the roots of q lie inside the unit disk U .

There is a probabilistic interpretation of this criterion in identification: If f is the transfer function of a P^2 stable system driven by a white noise δ , the output $y = f\delta$ is a stationary process. If the latter is to be modelled by a rational function p/q of order at most n , and if we put $\hat{y} = p/q\delta$, the minimization of the covariance of $y - \hat{y}$ is achieved when $\|f - p/q\|_2$ is itself minimal.

In this paper, we present an algorithm to find local best approximants of a given order n , which proceeds recursively by numerically solving differential equations over a compact subset of \mathbb{R}^n . This procedure is the first one, to our knowledge, for which convergence is guaranteed, at least generically. We then present a convincing experiment, and we finally list some open questions. Most proofs are just sketched, since our main concern here is not technical but rather to describe the procedure.

Let us first list a few known results concerning this question. It can be proved [see e.g. Baratchart (1986), Ruckebusch (1978) or Walsh (1962)] that the problem stated above has a solution. This solution is not always unique, but generically it is (Baratchart, 1987), though there might be lots of local minima. One can show that if f is not a rational function of order less than n , a case which will be implicitly ruled out in what follows, no local minimum can be of order less than n (Ruckebusch, 1978). In other words, one should always take advantage of all parameters at hand.

This observation leads to the conclusion that it is enough to minimize the norm over the set of irreducible fractions of H_2 of order exactly n . Since this set is a manifold, it is possible to use classical tools from optimization, like steepest descent algorithms. However, due to the shape of the gradient vector field and to the non-compactness of the domain over which we optimize, these methods may fail to converge. In this paper, we develop a different approach, which is based on the elimination of some parameters and gives rise to a much nicer geometric picture.

2. The function Ψ_n

Let P_n be the set of real polynomials of degree at most n , and \mathcal{P}_n^1 the subset of monic polynomials of degree n whose roots are in the disk U_r of radius r .

We look for

$$\min_{p \in P_{n-1}, q \in \mathcal{P}_n^1} \left\| \frac{p}{q} - f \right\|_2^2, \tag{3}$$

where $p \in P_{n-1}$ and $q \in \mathcal{P}_n^1$. Consider the n -dimensional linear subspace of H_2 defined by $V_q = P_{n-1}/q$. For fixed q , the minimum in (3) is obtained when p/q is the orthogonal projection $\pi_q(f)$ of f onto V_q . If we define a polynomial $L_n(q) \in P_{n-1}$ by the formula $L_n(q) = q\pi_q(f)$, we are thus led to minimize the function

$$\Psi_n: \mathcal{P}_n^1 \rightarrow \mathbb{R} \text{ defined by } \Psi_n(q) = \left\| f - \frac{L_n(q)}{q} \right\|_2^2.$$

The polynomial $L_n(q)$ must satisfy by definition

$$\forall j = 0, \dots, n-1, \left\langle f - \frac{L_n(q)}{q}, \frac{z^j}{q} \right\rangle = 0,$$

that is to say

$$\frac{1}{2i\pi} \int_{\Gamma} \left(f\left(\frac{1}{z}\right) - \frac{L_n(q)}{q} \left(\frac{1}{z}\right) \right) \frac{u}{q} (z) \frac{dz}{z} = 0,$$

for any complex polynomial u of degree at most $n-1$.

Let us define the function g holomorphic in U by putting

$$g(z) = f(1/z)/z. \tag{4}$$

Define further $\tilde{q} \in P_n$ as $z^n q(1/z)$. The roots, possibly at infinity, of \tilde{q} are the inverses of those of q . Similarly, we shall also denote by $\tilde{L}_n(q)$ the polynomial $z^{n-1} L_n(q)(1/z)$. With these notations, our integral equation becomes

$$\frac{1}{2i\pi} \int_{\Gamma} \left(g - \frac{\tilde{L}_n(q)}{\tilde{q}} \right) \frac{u}{q} dz = 0,$$

whenever u is a complex polynomial of degree at most $n-1$. Since $g - \tilde{L}_n(q)/\tilde{q}$ belongs to the Hardy space $H_2(U)$ of the unit disk U (Duren, 1970), the residue theorem applies giving that the above is satisfied if and only if $g - \tilde{L}_n(q)/\tilde{q}$ matches zero at each root of \tilde{q} , counting multiplicities, namely q should divide $g\tilde{q} - \tilde{L}_n(q)$. Thus, $\tilde{L}_n(q)$ is the unique polynomial of degree at most $n-1$ interpolating $g\tilde{q}$ at the zeroes of \tilde{q} , that is to say:

Proposition 1. The polynomial $\tilde{L}_n(q)$ is the remainder of the division of $g\tilde{q}$ by \tilde{q} .

A well-known integral representation for our remainder (Walsh, 1962) is

$$\tilde{L}_n(q) = \frac{1}{2i\pi} \int_{\Gamma} \frac{\tilde{q}g(\xi)}{q(\xi)} \left[\frac{q(\xi) - q(z)}{\xi - z} \right] d\xi, \tag{5}$$

where Γ is any contour contained in the domain of holomorphy of g that encompasses the zeroes of q . As usual, the independence of the integral from the contour follows from Cauchy theorem.

3. Extension of the domain of Ψ_n : smoothness

A monic polynomial of degree n , $q(z) = z^n + a_{n-1}z^{n-1} + \dots + a_0$ can be identified with the vector $(a_{n-1}, a_{n-2}, \dots, a_0)$ of \mathbb{R}^n , and this allows for \mathcal{P}_n^1 to be considered as an open subset of \mathbb{R}^n .

So far, Ψ_n has been defined only on \mathcal{P}_n^1 . Now, if we assume that f is holomorphic not only for $|z| > 1$ but also in a neighborhood of the unit circle T , we shall be able to extend Ψ_n to a smooth function defined on a neighborhood of the closure, in \mathbb{R}^n , of \mathcal{P}_n^1 . This closure will be denoted by Δ_n , and clearly consists of all real monic polynomials of degree n whose roots are in the closed unit disk \bar{U} . Our hypothesis on f , which will be assumed hereafter, is equivalent to requiring that g be analytic in the disk U_r of radius r for some $r > 1$ that will remain fixed in the sequel.

Proposition 2. The map Ψ_n extends to a smooth function $\Psi_n: \Delta_n \rightarrow \mathbb{R}$.

Proof. If $q \in \mathcal{P}_n^1$, the properties of the orthogonal projection show that

$$\Psi_n(q) = \left\| f - \frac{L_n(q)}{q} \right\|_2^2 = (f, f) - \left\langle f, \frac{L_n(q)}{q} \right\rangle. \tag{6}$$

Since the contour Γ may be deformed within the domain of

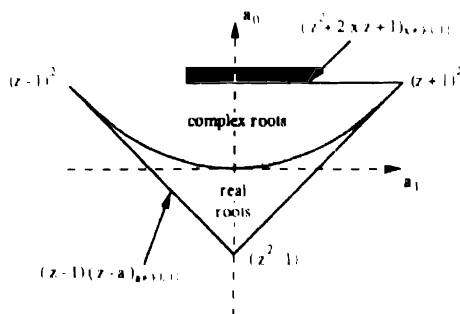


FIG. 1. The set Δ_2 of points $(a_1, a_0) \in \mathbb{R}^2$ such that the polynomial $z^2 + a_1z + a_0$ has all its roots in the closed unit disk.

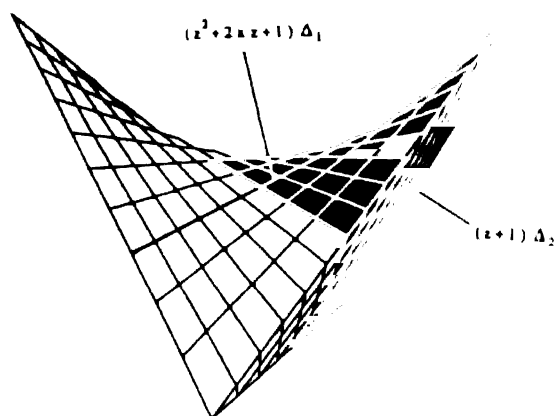


FIG. 2. A representation of Δ_1 and its boundary

$$\partial\Delta_1 = (z+1)\Delta_2 \cup (z-1)\Delta_2 \cup_{x \in (-1,1)} (z^2 + 2xz + 1)\Delta_1.$$

holomorphy of g so as to surround any n -tuple of points in U_r , the integral representation (5) obviously yields a smooth extension of L_n to a map $\mathcal{P}_n \rightarrow P_{n-1}$. Note that $\bar{L}_n(q)$ is still the remainder of the division

$$g\bar{q} = vq + \bar{L}_n(q). \quad (7)$$

Having this at our disposal, it is now sufficient by (6) to extend smoothly for every k , $q \rightarrow \langle f, z^k/q \rangle$ to \mathcal{P}_n . This can be done similarly by putting

$$\left\langle f, \frac{z^k}{q} \right\rangle = \frac{1}{2i\pi} \int_1^{\infty} g(\xi) \frac{z^k}{q(\xi)} d\xi,$$

which again allows for q to lie in \mathcal{P}_n . Q.E.D.

We now turn to Δ_n . It is plain to see that Δ_0 consists solely of the point 0, and Δ_1 of the segment $[-1, 1]$. It is easy to check that Δ_2 is the triangle with vertices $(-2, 1)$, $(2, 1)$ and $(0, -1)$ (see Fig. 1). It is more difficult to see what Δ_k looks like (see Fig. 2). In general, Δ_n is topologically a ball, as is proved in Baratchart and Olivi (1988), and we shall concentrate here on its boundary $\partial\Delta_n$, which is homeomorphic to a sphere. This set consists of polynomials in Δ_n having at least one root of modulus 1. Among such polynomials, one can distinguish between those having ± 1 as a root and those having no real root of modulus 1, but which are divisible by some polynomial of the form $z^2 + 2xz + 1$ with $x \in (-1, 1)$. This means that $\partial\Delta_n$ is the union of $(z-1)\Delta_{n-1}$, $(z+1)\Delta_{n-1}$, and $(z^2 + 2xz + 1)\Delta_{n-2}$ for $x \in (-1, 1)$. In other words, $\partial\Delta_n$ is made from two copies of Δ_{n-1} and infinitely many copies of Δ_{n-2} . As is already apparent when n equals 2 or 3, the boundary $\partial\Delta_n$ is not smooth, namely there are corners. The smooth part of $\partial\Delta_n$ may be characterized as the set of polynomials in Δ_n having exactly one irreducible factor over \mathbb{R} with roots of modulus one. Alternatively, they are interior points of the copies of Δ_{n-1} and Δ_{n-2} introduced above. The crux of the matter is the following lemma describing the behaviour of Ψ_n on $\partial\Delta_n$.

Lemma 1. Let $q \in \partial\Delta_n$, and suppose $q = q_u q_i$, where q_u is monic of degree k and has all its roots of modulus 1 while q_i is interior to Δ_{n-k} . Then $\Psi_n(q) = \Psi_{n-k}(q_i)$.

Proof. From (6), it is sufficient to prove that $L_n(q) = q_u L_{n-k}(q_i)$. But, since inverse and conjugate agree on T , we have $\bar{q}_u = \pm q_u$ and the result follows from (7). Q.E.D.

Let us denote by $\nabla_n(q)$ the gradient vector of Ψ_n at the point q . Later we will use the following consequence of Lemma 1.

Corollary. Let q_i as in the previous lemma, belong to the smooth part of $\partial\Delta_{n-k}$ and be such that q_i is a critical point of

Δ_{n-k} (note that $k = 1$ or 2). Then $\nabla_n(q)$ is orthogonal to $\partial\Delta_n$ and points outwards.

Proof. From Lemma 1, we see that the projection of $\nabla_n(q)$ on $\partial\Delta_n$ is just $\nabla_{n-k}(q_i)$, so that $\nabla_n(q)$ is orthogonal to $\partial\Delta_n$. Moreover, it cannot point inwards because this would imply that $L_{n-k}(q_i)/q_{n-k}$, which is rational of order $n-k$, is locally a best approximant to f among rational functions of order n , hence that f itself is rational of order $\leq n$. Q.E.D.

4. A generic algorithm to find a local minimum

As said before, the minimum value of Ψ_n on Δ_n can only be taken at some interior point of Δ_n since f is not rational of order less than n by hypothesis. As a consequence, such a point q must be a critical point of Ψ_n , and has to satisfy

$$\nabla_n(q) = 0.$$

We shall make two extra assumptions in what follows. First, we shall assume that ∇_k does not vanish on $\partial\Delta_k$ for $1 \leq k \leq n$. Second, k given as above, we shall ask all critical points of Ψ_k in Δ_k to be nondegenerate, i.e. to have a second derivative that is a nondegenerate quadratic form. These two properties hold generically, that is for almost every f in some sense, and we refer the reader to Baratchart and Olivi (1988) for a precise statement. They ensure in particular that critical points in Δ_k are finite in number.

Taking this for granted, we are now able to describe a procedure to determine a local minimum of Ψ_n . We first define one more bit of notation: if $q \in \Delta_k$ for some $k \leq n$, we define $\Psi(q)$ to be simply $\Psi_k(q)$. This new notation enables us to compare $q \in \Delta_k$ and $q' \in \Delta_k$, but we shall still use Ψ_k when referring to the behaviour on Δ_k only.

The algorithm proceeds as follows. Choose some interior point q_0 of Δ_n as an initial condition, and integrate the vector field $-\nabla_n$. There are two possibilities: Either we reach a critical point or we reach $\partial\Delta_n$. In the former case, if the critical point is a local minimum, we are done. Otherwise, since it is nondegenerate, the critical point will be unstable under small perturbations, thereby allowing us to continue the procedure. Since Ψ_n decreases, we cannot meet the same critical point twice, and because such points are finite in number, we eventually succeed or we reach $\partial\Delta_n$.

If we meet $q_k \in \partial\Delta_n$ (see Fig. 3), we decompose it as $q_k = q_u q_i$, where q_u has all its roots of modulus 1 and q_i has none. From Lemma 1, we see that $\Psi(q) = \Psi(q_i)$. Moreover, the degree k of q_i is nonzero, for Ψ_0 is a constant function whose value $\|f\|_2^2$ is the maximum of Ψ_n on Δ_n .

Now, q_i lies in the interior of Δ_k , so we can begin all over again with n replaced by k and q_0 by q_i . Since k decreases but remains strictly positive, we are bound to reach a local minimum of Ψ_m for some m satisfying $1 \leq m \leq n$, at some interior point q_m of Δ_m .

Consider now the point $q_1 = q_m(z+1)$ of $\partial\Delta_{m+1}$. It lies in a smooth region of the boundary, so that $-\nabla_{m+1}(q_1)$ is orthogonal to $\partial\Delta_{m+1}$ and points inwards by the corollary to Lemma 1. Therefore, integrating $-\nabla_{m+1}$ starting from q_1 (see Fig. 3) leads us to penetrate into the interior of Δ_{m+1} , so that the whole process can be carried over again, with n replaced by $m+1$. Since Ψ decreases continuously, we never meet twice the same critical point of Ψ_k for $1 \leq k \leq n$, and this ensures that the procedure eventually comes to an end. An end means precisely that we have reached the desired local minimum of Ψ_n .

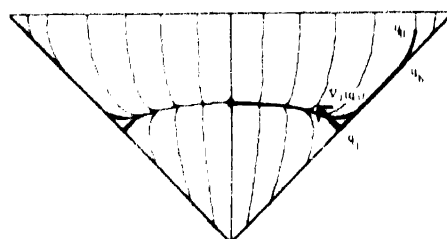


FIG. 3. Sketch of the algorithm for $n=2$ and $f(z) = -4z^{-1} + z^{-1}$.

5. The continuous time case

So far, we have only dealt with rational approximation in the Hardy space of the disk, as related to the identification of discrete-time L^2 -stable transfer functions. In practice, however, continuous-time systems mainly occur. A treatment similar to that in Section 1 could be given, but we just indicate briefly how the above technique applies. Let us call a continuous-time system L^2 -stable if its impulse response is in $L^2[0, \infty]$. The Paley-Wiener theorem (see e.g. Duren, 1970) asserts that the Laplace transform is an isometry between $L^2[0, \infty]$ and the real Hardy space \mathcal{H}_2 of right-half plane consisting of functions F holomorphic for $\Re(z) > 0$, satisfying the realness condition $F(\bar{z}) = \overline{F(z)}$, and such that

$$\sup_{y>0} \int_{-\infty}^{\infty} |F(x+iy)|^2 dy < \infty. \tag{8}$$

Here, the squared norm $\|F\|_{\mathcal{H}_2}^2$ is the supremum in (8) by definition. In other words, the set of transfer functions of L^2 -stable systems is precisely \mathcal{H}_2 . The rational approximation problem in this context consists in looking for

$$\min_{p/q} \left\| F - \frac{p}{q} \right\|_{\mathcal{H}_2}$$

where p/q ranges over all rational functions of order $\leq n$ in \mathcal{H}_2 . Note that such fractions are exactly transfer functions of stable systems of order at most n .

A possible interpretation of the L^2 criterion is as follows. Assume an L^2 system is driven by a white noise, so that the output is a stationary stochastic process

$$\xi(t) = \int_{-\infty}^t h(t-\tau) dW(\tau)$$

where W is a Wiener process. The variance of ξ , which is independent of t , is equal to $\|h\|_2^2$ (Doob, 1953). If h_n is the impulse response of a system of order at most n , and p/q its transfer function, and if we put

$$\xi_n(t) = \int_{-\infty}^t h_n(t-\tau) dW(\tau),$$

the variance of $\xi - \xi_n$ is $\|h - h_n\|_2^2$, which is also equal to $\|F - p/q\|_{\mathcal{H}_2}^2$, where F and p/q are respectively the Laplace transforms of h and h_n . Therefore, if we want a model of order at most n for the process, we minimize the covariance of the error by solving the above rational approximation problem.

Now, the question comes back to the one investigated

above. Indeed, if we put

$$\varphi: z \rightarrow \frac{z+1}{z-1} \quad \text{and} \quad \Phi(F)(z) = 2\sqrt{\pi} \frac{F \circ \varphi(z)}{z-1},$$

it turns out that Φ is an isometry between \mathcal{H}_2 and H_2 , and it is plain to see that Φ preserves rational functions and their order. Therefore, putting $f = \Phi(F)$ brings us back to the discrete-time case.

6. Numerical examples

Until now, we have assumed that the transfer function of the system under consideration exists and is perfectly known to us. But in practice, of course, this is never so. The system certainly exists, but the transfer function may not be defined. And even if it does, it is only known to us through a finite number of experiments, whereas the function f to be approximated in H_2 has to be completely defined if we actually want to run the algorithm. In fact, the definition of f from a set of experimental data is an arduous problem that was entirely bypassed in the above development.

For instance, if the transfer function F of a continuous-time system is computed through frequency response experiments, we are given its value at a certain number of points of the imaginary axis, so that f is only known at a finite number of points on the unit circle. To estimate its Fourier coefficients, the best we can hope for, in general, is to have a convergent procedure to estimate $\Phi(F)$ when the number of experiments increases. This is in the style of Baratchart (1989); however, we shall not go into further details here. We shall instead present examples where this step has been carried out by *ad hoc* methods. The procedure has been implemented on a computer, using a standard package for the numerical integration of ordinary differential equations. We chose the b.d.f. method. The computation of the gradient is done from explicit division formulas, which we do not derive here due to limited space, and can be carried out using the Euclidean algorithm since g is a polynomial in practice. This also implies that Ψ_n exists and is smooth on \mathbb{R}^2 . All along the integration, a control is made on the points of the path to verify that they lie in Δ_n by computing the roots of each corresponding polynomial. If we go outside Δ_n , we use dichotomy on the stepsize, to determine accurately the crossing point on the boundary. This gives the initial condition of a lower order integration as described in Section 4.

Several sets of data have been treated, most of them obtained from experimental measurements of an aircraft's high-frequency modes. Figures 4 and 5 show an example

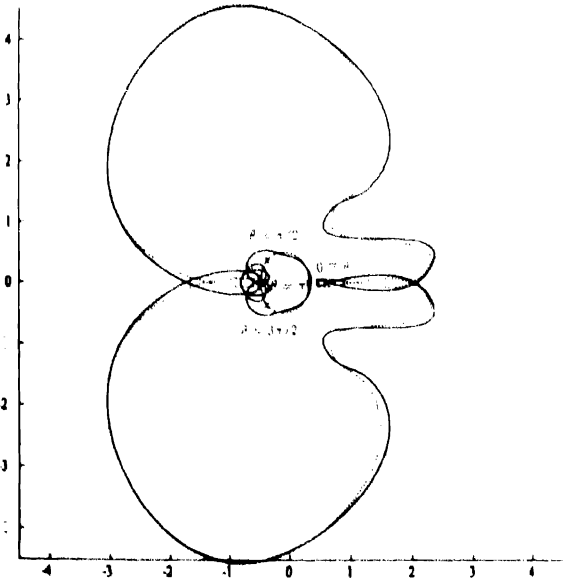


FIG. 4. The transfer function f (continuous line) and its best approximation of order 7 (dotted line).

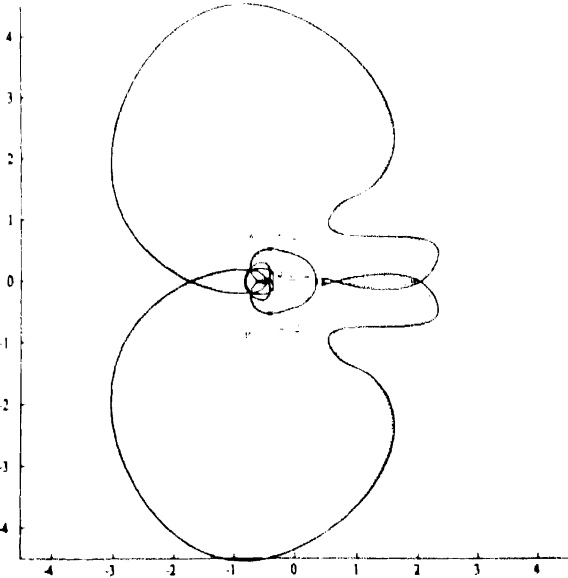


FIG. 5. The transfer function f (continuous line) and its best approximation of order 10 (dotted line).

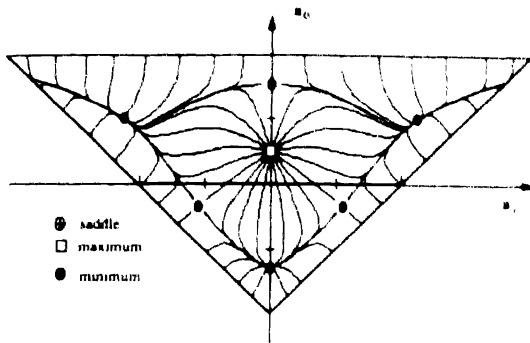


FIG. 6. Integral curves of the vector field ∇_z for $f(z) = 1/4z^{-1} + z^{-1}$.

selected as being difficult to approximate, where f is described through an estimation of its first 200 Fourier coefficients. Since the maximum error occurs when $|z| = 1$, we choose to represent the functions by the 2-dimensional plots of their values around the unit disk. The function f (continuous line) and its best approximant (dotted line) of order 7 are plotted on Fig. 4. The computation time on a SUN station 4/110 was 13 minutes. The result is not completely satisfactory and one has to go up to order 10 to get a better approximation. This time the computation time was 17 minutes. Figure 5 shows the corresponding approximation.

7. Conclusions

At this point, it is only fair to say that the procedure described above does not quite answer the original question, since it only ensures that we meet a *local* minimum, and not necessarily a *global* one. This deficiency is puzzling for there may be lots of local minima. Figure 6 shows already 3 local minima, when n is only equal to 2 and f is the simple rational function of order 3: $1/4z^{-1} + z^{-1}$.

Further investigations on this problem may be envisaged from two different viewpoints. On one hand, it is possible to consider this difficulty as intrinsic and cope with it, trying to find the global minimum at any cost. One may think of initializing the algorithm at enough points of the compact set Δ_n , to reach all local minima and compare between them. But the precise meaning of the word "enough" depends, of course, on f and n , and we are not able to give an *a priori* bound for it. Consequently, more efficient strategies should be investigated. For instance, we can restrict ourselves to initial points lying on $\partial\Delta_n$ provided n is large enough (Baratchart *et al.*, 1990). In examples we have met so far, initiating the algorithm from local minima on $\partial\Delta_n$ is in fact enough to exhaust the set of local minima in Δ_n . Since local minima on $\partial\Delta_n$ are solutions of the corresponding problem in Δ_{n-1} and Δ_{n-2} , thanks to lemma 1, this allows one to proceed recursively. We do not know, however, whether this property holds in some generality.

On the other hand, the fact that there are several local minima may cause discrepancy in identification. For instance, there are situations (Ruckebusch, 1978) when two distinct

rational functions of order n , say r_1 and r_2 , are both best approximants to f . Though these situations are exceptional (Baratchart, 1987), the L^2 identification problem at order n is not well-posed in the neighborhood of such an f , because juggling it slightly yields a best approximant which is close to r_1 or r_2 alternatively. Whether such a phenomenon is due to f or rather a consequence of inappropriate a value for n is not yet clear. This suggests that the physical meaning of the difficulties explained above should be further analysed. In particular, it would be of interest to derive conditions on f ensuring there are no local minima except the global one, at least for n sufficiently large. A subclass of Stieltjes functions, for instance, has been shown recently to have this property (Baratchart *et al.*, unpublished data), but a lot of work remains to be done in this direction.

A related problem is the behaviour of Ψ_n as $n \rightarrow \infty$. For instance, it is possible to prove (Baratchart *et al.*, 1990) that all critical points of Ψ_n converge to f in H_2 as $n \rightarrow \infty$. But the rate of convergence is likely to depend on the nature of the points (saddles or minima), and such questions are wide open.

Finally, in order to apply in a meaningful way to system theory, this technique has to be extended to the multi-input multi-output case, a question which is not yet settled. The main difficulty is to find an analogue to Ψ_n . This generalization is currently under investigation.

As a conclusion, let us express our point of view that rational approximation has something to offer in system theory and that differential tools are useful in smooth situations like the one arising here. The above algorithm is intended to be a modest contribution to this range of ideas.

References

- Baratchart, L. (1986). Existence and generic properties of L^2 approximants for linear systems. *Math. Control Inform.*, **3**, 89–101.
- Baratchart, L. (1987). Recent and new results in rational L^2 approximation. In R. F. Curtain (Ed.), *Modelling, Robustness and Sensitivity Reduction in Control Systems*. Springer, Berlin.
- Baratchart, L. (1990). Interpolation and Fourier coefficients in the Hardy space H_2 . In Kaashoek, M. A., J. H. van Schuppen and A. C. M. Ran (Eds.), *Realization and Modelling in System Theory—Proc. Int. Symp. MTNS-89*, Vol. 1, pp. 387–394. Birkhauser, Boston.
- Baratchart, L. and M. Olivi (1988). Index of critical points in rational L^2 approximation. *Syst. Control Lett.* **10**, 163–174.
- Baratchart, L., M. Olivi and F. Wielonsky. (1990). Asymptotic properties of critical points in rational L^2 approximation. *9th Conf. Analysis and Optimization of Systems*, Antibes, to appear.
- Doob, J. L. (1953). *Stochastic Processes*. Wiley, New York.
- Duren, P. (1970). *Theory of H^p Spaces*. Academic Press, New York.
- Fuhrmann, P. A. (1981). *Linear Systems and Operators in Hilbert Space*. McGraw-Hill, New York.
- Ruckebusch, G. (1978). Sur l'approximation rationnelle des filtres, Rapport No 35 CMA Ecole Polytechnique.
- Walsh, J. L. (1962) *Interpolation and Approximation by Rational Functions in the Complex Domain*, A.M.S. Publications, vol. 10.

Technical Communique

A New Result on Relative Gain Array, Niederlinski Index and Decentralized Stability Condition: 2×2 Plant Cases*

MIN-SEN CHIU† and YAMAN ARKUN‡

Key Words—Decentralized stability; relative gain array; Niederlinski Index.

Abstract—This note discusses the decentralized stability condition for a 2-channel decentralized control system where each channel is a single-input-single-output (SISO) control loop. Stabilization of each individual loop (with the other loop open) is not required, which had been the case in previous work on the Relative Gain Array (RGA) and the Niederlinski Index (NI). Therefore this new result is a generalization of previous RGA and NI stabilization conditions.

1. Introduction

FOR A 2-channel decentralized control system where each channel is a single-input-single-output (SISO) control loop, decentralized stability condition can be related to the steady-state gain of the open loop system using the Relative Gain Array (RGA) and the Niederlinski Index (NI) (Grosdidier *et al.*, 1985; Niederlinski, 1971). Both of these quantities have found wide applications in process control in particular. In the derivation of these results (e.g. see Grosdidier *et al.*, 1985) the stability of the individual loop (with the other loop open) is assumed. Although such an assumption can be justified based on practical grounds like reliability, it is not a general theoretical result. Therefore in this work this assumption is relaxed and new stability conditions related to RGA and NI are obtained.

2. Preliminaries

Assumption 1 holds in the following development.

Assumption 1:

- The plant is described by $G(s) = [g_{ij}(s)]$ for $i, j = 1, 2$ and $G(s)$ is stable, strictly proper or semi-proper.
- Decentralized controller

$$C(s) = \begin{bmatrix} c_1(s) \\ c_2(s) \end{bmatrix}$$

has two SISO loops and each loop contains integral action.

Under Assumption 1 the decentralized feedback control scheme is shown in Fig. 1. The corresponding decentralized internal model control (IMC) structure is given in Fig. 2 with the following equivalence:

$$\begin{aligned} G_m(s) &= \text{diag}[g_{ii}(s)] \quad \forall i = 1, 2 \\ Q(s) &= \text{diag}[q_i(s)] \quad \forall i = 1, 2 \end{aligned} \quad (1)$$

* Received 5 February 1990; revised 6 August 1990; received in final form 10 September 1990. Recommended for publication by Editor W. S. Levine. An earlier version of this paper was presented at the BILCON Conference, Ankara, Turkey in July 1990.

† School of Chemical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0100, U.S.A.

‡ To whom all correspondence should be addressed.

and

$$q_i(s) = \frac{c_i(s)}{s + g_{ii}(s)c_i(s)} \quad \forall i = 1, 2 \quad (2)$$

where $q_i(s)$ is allowed to be *unstable* which implies loop i ($i = 1, 2$) to be *unstable* by itself. It is clear from (2) that $q_i(s)$ has to satisfy

$$q_i(0) = \frac{1}{g_{ii}(0)} = \frac{1}{g_{ii}^0}$$

to ensure the integral action of both loops. Later we make use of this equivalence between the decentralized feedback and IMC structure to prove the results.

The Relative Gain Array (RGA) was developed by Bristol (1966). Given the steady-state gain of $G(s)$, i.e. $G(0) = [g_{ij}^0]$ for $i, j = 1, 2$ where g_{ij}^0 is real, RGA is defined as

$$\Lambda = [\lambda_{ij}] \quad \forall i, j = 1, 2 \quad (3)$$

where

$$\lambda_{11} = \lambda_{22} = \frac{g_{11}^0 g_{22}^0}{g_{11}^0 g_{22}^0 - g_{12}^0 g_{21}^0} \quad (4)$$

$$\lambda_{12} = \lambda_{21} = 1 - \lambda_{11} = 1 - \lambda_{22} \quad (5)$$

In the 2×2 plant cases, the Niederlinski Index (NI) (Niederlinski, 1971) is defined to be

$$NI = \frac{\det[G(0)]}{\prod_{i=1}^2 g_{ii}^0} = \frac{g_{11}^0 g_{22}^0 - g_{12}^0 g_{21}^0}{g_{11}^0 g_{22}^0} = \frac{1}{\lambda_{11}} \quad (6)$$

Under the assumption that the individual loop (i.e. loop 1 or 2) is *stable* when the other loop is open (i.e. loop 2 or 1), it is proved that $\lambda_{ii} > 0$ (or $NI > 0$) is the necessary and sufficient condition for decentralized stability (Grosdidier *et al.*, 1985). However, *a priori* stabilization of individual loops is not a *prerequisite* for the stability of the decentralized system. Therefore in this work this requirement is relaxed and the following questions are answered: Is there any restriction on the number of unstable closed-loop poles of loops 1 and 2? If yes, how many unstable closed-loop poles can one assign to loops 1 and 2?

3. Stability conditions

The next proposition states the decentralized stability condition for a 2×2 plant. Denote

$$\Delta(s) = 1 - g_{12}(s)q_2(s)g_{21}(s)q_1(s) \quad (7)$$

Proposition 1. Let Assumption 1 hold. The necessary and sufficient condition for decentralized stability is that $\Delta(s)$ does not contain any RHP-zeros.

Proof. Let $\hat{Q}(s)$ be the Youla Parametrization as defined in Youla *et al.* (1976).

$$\hat{Q}(s) = [\hat{q}_{ij}(s)] \quad \forall i, j = 1, 2$$

where $\hat{q}_{ij}(s)$ s (for $i, j = 1, 2$) are *stable* semi-proper or strictly proper rational functions.

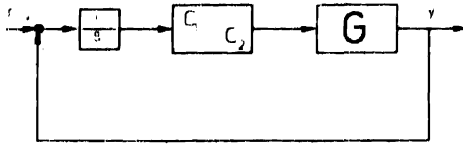


FIG. 1. Decentralized feedback control scheme.

Under Assumption 1, the decentralized stability conditions are (Manousiouthakis, 1989)

$$\bar{q}_{12}(s)g_{21}(s) = \bar{q}_{21}(s)g_{12}(s) \quad (8)$$

$$\bar{q}_{11}(s)\bar{q}_{22}(s) - \bar{q}_{12}(s)\bar{q}_{21}(s) = \frac{-\bar{q}_{12}(s)}{g_{12}(s)} = \frac{-\bar{q}_{21}(s)}{g_{21}(s)} \quad (9)$$

The decentralized IMC controllers in this note and the decentralized feedback controllers in Manousiouthakis (1989) are related by the following equality

$$Q(s)(I - G_m(s)Q(s))^{-1} = (I - \bar{Q}(s)G(s))^{-1}\bar{Q}(s) \quad (10)$$

It is noted that the decentralized stability constraints on $\bar{q}_i(s)$ [(8), (9)] will render the right hand side of (10) a diagonal matrix.

After some algebra (10) yields

$$q_1(s) = \frac{\bar{q}_{11}(s)}{1 - \bar{q}_{12}(s)g_{21}(s)} = \frac{-\bar{q}_{21}(s)}{\bar{q}_{22}(s)g_{12}(s)} \quad (11)$$

$$q_2(s) = \frac{\bar{q}_{22}(s)}{1 - \bar{q}_{21}(s)g_{12}(s)} = \frac{-\bar{q}_{12}(s)}{\bar{q}_{11}(s)g_{21}(s)} \quad (12)$$

Therefore, from (7), (8), (9), (11), (12)

$$\Delta(s) = \frac{1}{1 - \bar{q}_{21}(s)g_{12}(s)} \quad (13)$$

$$= \frac{\text{den}(\bar{q}_{21}(s)g_{12}(s))}{\text{den}(\bar{q}_{21}(s)g_{12}(s)) - \text{num}(\bar{q}_{21}(s)g_{12}(s))} \quad (14)$$

where num(\cdot) and den(\cdot) denote the numerator and denominator of (\cdot) respectively. Since both $\bar{q}_{21}(s)$ and $g_{12}(s)$ are stable rational transfer functions, it can be concluded that $\Delta(s)$ does not contain any RHP-zeros and the proof is complete. \square

The next lemma relates the stability conditions of a 2×2 decentralized control system to RGA and NI

Lemma 1. If Assumption 1 holds and

1. λ_u (or NI) < 0 , then one of the following conditions holds:

- (a) the loops 1 and 2 and the system are unstable;
- (b) the loops 1 and 2 are stable, but the system is unstable;
- (c) only one of the loops 1 and 2 is stable, but the system is unstable;
- (d) only one of the loops 1 and 2 is stable, but the system is stable.

2. λ_u (or NI) > 0 , then one of the following conditions holds: (1a), (1b), (1c) and

- (a) the loops 1 and 2 and the system are stable;
- (b) the loops 1 and 2 are unstable, but the system is stable.

Proof. We only show the proof of part 1, since the same reasoning follows directly to prove part 2. From Grosdidier

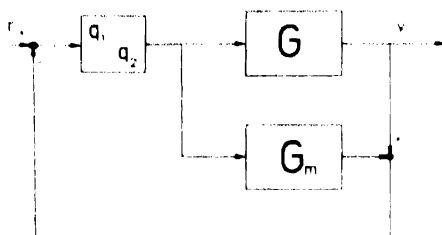


FIG. 2. Decentralized internal model control scheme.

et al. (1985), one of the following holds when λ_u (or NI) < 0 :

- (i) $h_1(0) < 0$, $h_2(0) < 0$, $\det(H(0)) < 0$;
- (ii) $h_1(0) > 0$, $h_2(0) > 0$, $\det(H(0)) < 0$;
- (iii) either one of $h_1(0)$ and $h_2(0)$ is negative and $\det(H(0)) > 0$

where $h_i(s) = g_u(s)c_i(s)$ for $i = 1, 2$ and $H(s) = G(s) \text{diag}[c_i(s)]_{i=1,2}$.

Since $\det(H(0)) > 0$ is only a *necessary* condition for decentralized stability for 2×2 and larger systems while $h_i(0) > 0$ is a *necessary and sufficient* one for the stability of the individual loops, (1a) follows from (1), (1b) from (ii), (1c), (1d) from (iii). \square

The next proposition follows directly from Lemma 1.

Proposition 2. If Assumption 1 holds and the stability of the decentralized control system is required, then one of the following conditions holds:

1. λ_u (or NI) < 0 , one of the loops 1 and 2 has to be unstable;
2. λ_u (or NI) > 0 , both loops 1 and 2 have to be unstable;
3. λ_u (or NI) > 0 , both loops 1 and 2 have to be stable.

It is noted that (3) of Proposition 2 was discussed in Grosdidier *et al.* (1985) and Niederlinski (1971), while (1), (2) of Proposition 2 remain open. This is not surprising. The assumption that the individual loop is stabilized in Grosdidier *et al.* (1985) and Niederlinski (1971) precludes the discussion of (1), (2) in Proposition 2. The next section gives the main result of this note.

4. Main result

Theorem 1. Let Assumption 1 hold. The necessary and sufficient condition for decentralized stability is either one of the following conditions:

- (1) If λ_u (or NI) < 0 , one of the two loops has only *one* unstable closed-loop pole, while the other loop is stable;
- (2) If λ_u (or NI) > 0 , both loops have to be stable.

Proof. The proof is by contradiction.

(1) If λ_u (or NI) < 0 , one of the two loops has to be unstable from (1) of Proposition 2. Suppose that one can assign more than *one* unstable pole, say *two* unstable poles to loop 1 (the same reasoning applies in a straightforward fashion to more than two unstable poles cases):

$$q_1(s) = \frac{n_1(s)n_2(s)}{(s-p_1)(s-p_2)d_1(s)} \quad \text{Re}\{p_i\} > 0, \quad i = 1, 2$$

$q_2(s)$ to be any stable, proper, rational transfer function

where $\text{Re}\{\cdot\}$ denotes the real part of $\{\cdot\}$, $d_1(s)$ is a polynomial with only LHP-zeros, and $n_1(s)$, $n_2(s)$ are the polynomials such that $q_1(s)$ and $q_2'(s)$, $q_2''(s)$ defined below are proper. $q_1(s)$, $q_2(s)$ are designed such that $\Delta(s)$ of (7) satisfies Proposition 1. Nevertheless the following design also stabilizes the plant:

$$q_1'(s) = \frac{n_1(s)}{(s-p_1)d_1(s)}$$

$$q_2'(s) = \frac{n_2(s)}{s-p_2} \cdot q_2(s)$$

because $q_1(s)q_2(s) = q_1'(s)q_2'(s)$ and hence $\Delta(s) = \Delta'(s)$.

But then this *contradicts* (1) of Proposition 2. Therefore one has to destabilize any one loop with only *one* unstable closed-loop pole in order to stabilize the plant.

(2) If λ_u (or NI) > 0 , one possibility is that the two loops have to be unstable from (2) of Proposition 2. Suppose that k and m unstable poles are assigned to loops 1 and 2 respectively, i.e.

$$q_1''(s) = \frac{n_3(s)n_4(s)}{d_2(s) \prod_{i=1}^{k+2} (s-p_i)}$$

$$q_2''(s) = \frac{n_5(s)}{d_3(s) \prod_{i=1}^{k+m+2} (s-p_i)}$$

where $\text{Re}\{p_i\} > 0$ for $i = 3 \sim k + m + 2$ and $d_2(s)$, $d_3(s)$ are polynomials with only LHP-zeros. $n_1(s)$, $n_2(s)$ and $n_3(s)$ are the polynomials such that

$$\frac{n_1(s)}{\prod_{i=1}^{k+2} (s - p_i)}, \quad \frac{n_2(s)}{d_2(s)} \quad \text{and} \quad q_2^*(s)$$

are proper. $q_1^*(s)$, $q_2^*(s)$ are designed such that $\Delta^*(s)$ of (7) satisfies Proposition 1. But then the next design also stabilizes the plant

$$q_1^*(s) = \frac{n_1(s)}{d_2(s)}$$

$$q_2^*(s) = \frac{n_2(s)}{\prod_{i=1}^{k+2} (s - p_i)} \cdot q_2^*(s)$$

because $q_1^*(s)q_2^*(s) = q_1^*(s)q_2^*(s)$ and $\Delta^*(s) = \Delta^*(s)$. However this contradicts either (2) or (3) of Proposition 2. Hence (2) of Proposition 2 is false and one has to stabilize both loops for achieving decentralized stabilization. \square

The next example illustrates Theorem 1.

Example 1.

$$G(s) = \frac{1}{1 + 0.11s} \begin{bmatrix} 1 & 0.28 \\ 0.85 & 0.004 \end{bmatrix}$$

It is checked that $NI < 0$ and three different designs are discussed.

(1) Loops 1 and 2 are stable; for example, choose

$$q_1(s) = 1 \quad (15)$$

$$q_2(s) = 250 \frac{1 + 0.11s}{1 + \epsilon s} \quad \epsilon > 0 \quad (16)$$

Then the numerator of $\Delta(s)$ is

$$\text{num}(\Delta(s)) = \epsilon s^2 + (9.09\epsilon + 1)s - 531.818 \quad (17)$$

It is obvious that (17) has a RHP-zero for any positive ϵ and $G(s)$ can not be stabilized.

(2) Loop 1 is stable and loop 2 is unstable with two unstable poles. $q_1^*(s)$ is the same as (15) and

$$q_2^*(s) = 250 \frac{1 + 0.11s}{(-1 + \epsilon s)(-1 + \phi s)} \quad \epsilon, \phi > 0. \quad (18)$$

It is routine to calculate that

$$\text{num}(\Delta^*(s)) = 0.11\epsilon\phi s^4 + (\epsilon\phi - 0.11\epsilon - 0.11\phi)s^2 - (\epsilon + \phi - 0.11)s - 58.5. \quad (19)$$

Again, $G(s)$ can not be stabilized by any positive ϵ and ϕ

(3) Loop 1 is stable and loop 2 is unstable with one unstable pole. $q_1^*(s)$ is of (15) and

$$q_2^*(s) = -250 \frac{1 + 0.11s}{-1 + \epsilon s} \quad \epsilon > 0 \quad (20)$$

$$\text{num}(\Delta^*(s)) = 0.11\epsilon s^2 + (\epsilon - 0.11)s + 58.5. \quad (21)$$

The necessary and sufficient condition for (21) having LHP-zeros only is

$$\epsilon > 0.11. \quad (22)$$

It is interesting to see the physical interpretation of the stability constraint, (22). From (2), (20) the feedback controller of loop 2 is given by

$$\frac{1}{s} c_2^*(s) = -\frac{27.5}{\epsilon} \left(1 + \frac{1}{0.11s} \right) \quad (23)$$

Therefore the feedback gain is limited by (22). It should be tuned below $27.5/0.11 = 250$ to guarantee the stability of $G(s)$.

5. Conclusion

A new decentralized stability condition related to RGA and NI without requiring the stabilization of the individual loops is given. This result should be viewed as a generalization of previous work on RGA and NI. The result is limited to 2×2 plants under multi-loop SISO control and should hopefully motivate further development for larger systems.

Acknowledgements—The authors would like to thank Prof. M. Morari for his helpful comments on the parametrization of stabilizing controllers. The financial support of National Science Foundation is gratefully acknowledged.

References

1. Bristol, E. H. (1966). On a new measure of interaction for multivariable process control. *IEEE Trans. Aut. Control*, **AC-11**, 133–134.
2. Grosdidier, P., M. Morari and B. R. Holt (1985). Closed-loop properties from steady-state gain information. *Ind. Engng Chem. Fund.*, **24**, 221–235.
3. Manousiouthakis, V. (1989). On the parametrization of all decentralized stabilizing controllers. *Proc. ACC*, Pittsburgh, PA, pp. 2108–2111.
4. Niederlinski, A. (1971). A heuristic approach to the design of linear multivariable interacting control systems. *Automatica*, **7**, 691–701.
5. Youla, D. C., H. A. Jabr and J. J. Bongiorno (1976). Modern Wiener-Hopf design of optimal controllers—Part II. The multivariable case. *IEEE Trans. Aut. Control*, **AC-21**, 319–338.

Technical Communique

Passivity Properties for Stabilization of Cascaded Nonlinear Systems*

ROMEO ORTEGA†

Key Words—Nonlinear systems; stability; input-output stability

Abstract—We study the problem of global smooth stabilization of cascade compositions of nonlinear globally asymptotically stable systems. Our main result is that global stabilization can be achieved if we can render the first system strictly passive for an output which spans the “unstable part” of the vector field of the second system. This result generalizes earlier stabilizability conditions for the case when the first system is linear. Also, it provides some insight on the (“energy dissipation”) properties of the dependence of the second system vector field and the first systems “output” needed to achieve global stabilization.

1. Introduction

A PROBLEM of current interest in nonlinear control theory is the establishment of sufficient conditions for global smooth stabilization of cascaded nonlinear systems of the form (see Fig. 1):

$$\dot{x} = f(x, \xi), \quad x \in \mathbb{R}^n, \quad \xi \in \mathbb{R}^m \quad (1a)$$

$$\dot{\xi} = m(\xi) + G(\xi)u, \quad u \in \mathbb{R}^m \quad (1b)$$

where $f: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, $m: \mathbb{R}^m \rightarrow \mathbb{R}^m$ and $g: \mathbb{R}^m \rightarrow \mathbb{R}^m$ are smooth (i.e. of class C^∞) functions and assuming that

(A1) the system $\dot{x} = f(x, 0)$, $x \in \mathbb{R}^n$, has 0 as a globally asymptotically stable (GAS) equilibrium, or for short, is GAS; and

(A2) the system $\dot{\xi} = m(\xi)$ is also GAS.

It has been shown in Sontag (1988, 1989), that the cascade connection of two GAS systems does not yield, in general, a GAS system. However, if the stronger requirement of input to state stability [Definition 2.1 in Sontag (1989)] is imposed on the second system (1a), then the cascade is again GAS. A further refinement to this result was recently established in Seibert and Suarez (1990) where the weaker “small input bounded output” condition is shown to be sufficient for stability. On the other hand, motivated by the recent results on input-output linearization (see e.g. Isidori, 1989), Kokotovic and Sussman (1989) and Saberi *et al.* (1990) have studied the particular case of the problem above when the first system (1b) is a linear controllable system.

The construction of the stabilizing law in Kokotovic and Sussman (1989) is a variant of the cancellation procedure used in adaptive control (see e.g. Narendra and Annaswamy, 1989). This cancellation is possible if we can define an “output” $y = C\xi$ such that: (i) the transfer function $u \rightarrow y$ is strictly positive real [equivalently the map $u \rightarrow y$ is strictly passive, (Desoer and Vidyasagar, 1975)]; and (ii) The “unstable part” of the nonlinear system vector field belongs to the span of this output, i.e. if $f(x, \xi) - f(x, 0) \in \text{span}\{y_1, \dots, y_m\}$. A characterization of the passivity condition in terms of the transfer function leading Markov parameter and its stable invertibility is given in Saberi *et al.* (1990). This variant of the Kalman–Yakubovich–Popov (KYP) lemma allows the authors using dynamic compensators, to extend the result of Kokotovic and Sussman (1989) to a larger class of systems and to provide an interpretation of the required dependence of x and ξ in terms of the linear system zero dynamics.

In this paper we extend the result of Kokotovic and Sussman (1989) to the case when both systems are nonlinear. We follow the procedure of Kokotovic and Sussman (1989) described above to explore the “energy dissipation” properties of the dependence of f on ξ so as to insure stabilizability of the interconnected system. We show that global stabilization can be achieved if we can render the first system strictly passive for an output which spans the “unstable part” of the vector field of the second system.

2. Main result

The following definition of strict passivity will be used in the sequel.

Definition 1. Consider the state-space nonlinear system $\dot{x} = f(x) + G(x)u$, $x(0) = x_0 \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y = h(x) \in \mathbb{R}^m$. We say the mapping $H: u \rightarrow y$ is strictly passive relative to the functions $V(x)$ and $\beta(x)$ iff $V(\cdot): \mathbb{R}^n \rightarrow \mathbb{R}$ is a positive definite function of the state, $\beta(\cdot): \mathbb{R}^m \rightarrow \mathbb{R}$ is a class \mathcal{K} function and for all u, x, y solutions of the system equations and for any $t > 0$ we have

$$\int_0^t y^T u \, d\tau \leq V(x(t)) - V(x(0)) + \int_0^t \beta(|x|) \, d\tau.$$

Remark 1 We have included in our definition of strict passivity the functions $V(x)$ and $\beta(x)$. See also Hill and Moylan (1980) for a less restrictive, but similar in spirit, definition.

Remark 2 Notice that if $x = 0$ is an equilibrium then the strict passivity condition implies that $x = f(x, 0)$ is GAS. Conversely, if $x = f(x, 0)$ is GAS we can define an “output” such that the strict passivity condition is met, e.g. $y = \nabla_x V(x)^T G(x)$, with $V(x)$ a strict Lyapunov function for $x = f(x, 0)$.

We are in position to present our main result, whose proof is given in the next section.

Proposition 1. The cascade connection (1) with assumption (A1) is smoothly stabilizable if

(A2') The system (1b) defines a strictly passive operator $H_2: u \rightarrow y$ for an “output” $y = h(\xi): \mathbb{R}^m \rightarrow \mathbb{R}^m$ which satisfies the spanning condition

$$f(x, \xi) - f(x, 0) = \sum_{j=1}^m y_j f_j(x, \xi) \quad (2)$$

for some smooth functions $f_j: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, $j = 1, \dots, m$. ■

Remark 3 In view of Remark 2, our proposition shows that the cascade connection of two GAS systems can be made GAS if the “output”

$$y := \nabla_\xi V_2(\xi)^T G(\xi)$$

satisfies the spanning property (2), for some $V_2(\xi)$ -strict Lyapunov function of $\dot{\xi} = m(\xi)$.

Remark 4. Similarly to Kokotovic and Sussman (1989), stabilization is also achievable if the spanning assumption (2) is expressed as the existence of f_j , $j = 0, \dots, m$ so that

$$f(x, \xi) = f_0(x, \xi) + \sum_{j=1}^m y_j f_j(x, \xi).$$

Notice the dependence on ξ of f_0 . In this case, the condition of GAS of $\dot{x} = f(x, 0)$ should be replaced by global stability of $\dot{x} = f(x, \xi)$ for each fixed value of ξ . To complete the proof of the proposition an additional argument based on LaSalle's invariance principle is needed. For further details see Kokotovic and Sussman (1989).

* Received 27 June 1990; received in final form 29 September 1990. Recommended for publication by Editor H. Kwakernaak.

† National University of Mexico, DEPI-UNAM, P.O. Box 70-256, 04510 Mexico, D.F.

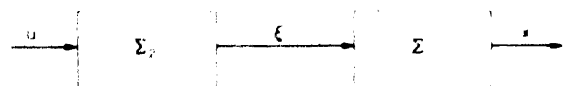


FIG. 1. Block diagram representation of (1) in terms of operators $\Sigma_1: \xi \rightarrow x$ for (1a) and $\Sigma_2: u \rightarrow \xi$ for (1b).

3. Proof of Proposition 1

To explain our approach we refer to the operator representation of (1) depicted in Fig. 1. Our proof is based on the idea of determining a feedback control $u(x, \xi)$ that will make (1) a feedback interconnection of two strictly passive operators. GAS of the closed-loop system will then follow from strict passivity of the feedback loop.

To this end, define first for system (1b) an operator $H_2: u \rightarrow y$ as

$$H_2: \begin{cases} \dot{\xi} = m(\xi) + G(\xi)u \\ y = h(\xi) \end{cases} \quad (3)$$

which, under assumption A2', is strictly passive, i.e. for all u, x, y solution of (3) and for all $t \geq 0$

$$\int_0^t y^1 u \, d\tau \geq V_2[\xi(t)] - V_2[\xi(0)] + \int_0^t \beta_2(|\xi|) \, d\tau \quad (4)$$

with V_2, β_2 positive definite and class \mathcal{K} functions respectively

Referring to Fig. 1, we notice that H_2 is the composition of Σ_2 and the output function $h(\cdot)$. The question is now whether we can decompose Σ_1 into the composition of $h(\cdot)$ and a strictly passive operator. If this is possible, we can, choosing the control equal to the output of the latter operator, obtain the desired feedback interconnection of two strictly passive operators.

The spanning condition (2) and the GAS assumption A1 allow this decomposition since, under this condition we can define for (1a) an operator $H_1: y \rightarrow z := [z_1, \dots, z_m]^T$

$$H_1: \begin{cases} \dot{x} = f(x, 0) + \sum_{j=1}^m y_j f_j(x, \xi) \\ z_j = \nabla_x V_1^1(x) f_j(x, \xi), \quad j = 1, 2, \dots, m \end{cases}$$

with $V_1(x)$ a strict Lyapunov function for $\dot{x} = f(x, 0)$. This operator can also be shown to be strictly passive. To this end, evaluate

$$\begin{aligned} \int_0^t y^1 z \, d\tau &= \sum_{j=1}^m \int_0^t y_j \nabla_x V_1^1(x) f_j \, d\tau = \int_0^t [V_1 - \nabla_x V_1^1 f(x, 0)] \, d\tau \\ &\geq V_1[x(t)] - V_1[x(0)] + \int_0^t \beta_1(|x|) \, d\tau \end{aligned} \quad (5)$$

where $\beta_1(\cdot)$ is a class \mathcal{K} function whose existence is guaranteed by the stability assumption A1.

Now, setting the control

$$u_j(x, \xi) = \nabla_x V_1^1(x) f_j(x, \xi) = z_j, \quad j = 1, \dots, m \quad (6)$$

we get a feedback interconnection of strictly passive operators (see Fig. 2). To prove that the overall system is GAS we combine (4), (5) and (6) to get

$$\begin{aligned} V_1[x(t)] + V_2[\xi(t)] &\leq V_1[x(0)] + V_2[\xi(0)] \\ &\quad - \int_0^t \beta_1(|x|) \, d\tau - \int_0^t \beta_2(|\xi|) \, d\tau \end{aligned}$$

which, invoking the arguments of Theorem 6 in Hill and Moylan (1980), completes the proof. ■

Remark 5. It is easy to show that the control law construction described above reduces, when the first system is linear, to the one proposed in Kokotovic and Sussman (1989) as follows. In this case (1b) is described by $\dot{\xi} = A\xi + Bu$. Now, choose a stabilizing control $u = K\xi$ and solve the Lyapunov equation $(A + BK)^T P + P(A + BK) = -Q, Q \succ 0$. The theorem insures stabilization if (1a) satisfies (2) with $y = B^T P\xi$.

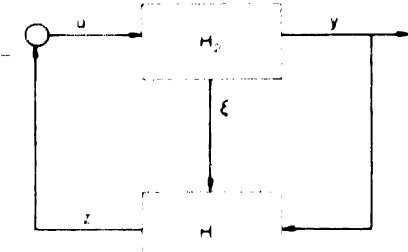


FIG. 2. Block diagram representation of the closed-loop system.

4. Example

In this section we apply our stabilization procedure to the following example

$$\dot{x} = x(x^2 \xi^2 - 1) \quad (7a)$$

$$\dot{\xi} = -\xi + \xi u. \quad (7b)$$

It can be shown that even though each independent system is GAS, e.g. $\dot{x} = -x$ and $\dot{\xi} = -\xi$, the cascade connection is not because (when $u = 0$) the set $x^2 y^2 = 1$ is invariant. Therefore, trajectories starting in the set $x^2 y^2 \geq 1$ cannot converge to zero. See also Seibert and Suarez (1990).

Following the design procedure above let us calculate $f(x, \xi) = f(x, 0)$ for (7a) which gives

$$f(x, \xi) = f(x, 0) = x^3 \xi^2.$$

The question is whether we can find an "output" $y = h(\xi)$ for (7b) such that the operator $H_2: u \rightarrow y$ is strictly passive and satisfies

$$x^3 \xi^2 = f_1(x, \xi) y$$

for some $f_1(x, \xi)$. To this end consider $V_2(\xi) := \xi^2/2$, whose derivative along the trajectories of (7b) yield

$$\dot{V}_2 = -\xi^2 + \xi^2 u.$$

Therefore, setting $y = \xi^2$ and $f_1(x, \xi) = x^3$ we attain both the spanning and strict passivity conditions. The stabilizing control is obtained from (6) as $u = -x^4$.

Acknowledgements—The author would like to express his gratitude to Professors P. Kokotovic, R. Suarez and E. Sontag for having provided him with early versions of their papers.

References

Desoer, C. and M. Vidyasagar (1975). *Feedback Systems: Input-Output Properties*. Academic Press, New York.

Hill, D. and P. Moylan (1980). Connections between finite-gain and asymptotic stability. *IEEE Trans. Aut. Control*, **AC-25**, 931–936.

Isidori, A. (1989). *Nonlinear Control Systems*, 2nd ed. Communications and Control Engineering Series. Springer, Berlin.

Kokotovic, P. and H. Sussman (1989). A positive real condition for global stabilization of nonlinear systems. *Syst. Control Lett.*, **13**, 125–133.

Narendra, K. and A. Annaswamy (1989). *Stable Adaptive Systems*. Prentice-Hall, Englewood Cliffs, New Jersey.

Saberi, A., P. Kokotovic and H. Sussmann (1990). Global stabilization of partially linear composite systems. *SIAM J. Control Optimiz.*, **28**, 1491–1506.

Seibert, P. and R. Suarez (1990). Global stabilization of nonlinear cascade systems. *Syst. Control Lett.*, **14**, 347–352.

Sontag, E. (1990). Further facts about input to state stabilization. *IEEE Trans. Aut. Control*, **AC-35**, 473–476.

Sontag, E. (1989). Smooth stabilization implies coprime factorization. *IEEE Trans. Aut. Control*, **AC-34**, 435–443.

Technical Communique

An Elementary Derivation of the Maximum Likelihood Estimator of the Covariance Matrix, and an Illustrative Determinant Inequality*

SEPPO KARRILA†‡ and TAPPIO WESTERLUND†

Key Words—Maximum likelihood estimation; estimation; determinants, optimization, least-squares estimation.

Abstract—The unique maximum likelihood estimate of the covariance matrix of normally distributed random vectors is derived by use of elementary linear algebra leading to simple scalar equations. In addition the application of a determinant inequality, also derived here, shows that a standard "derivation" of the maximum likelihood estimate is fallacious.

Introduction

IN SOME textbooks on estimation theory [for example Goodwin and Payne (1977), p. 48] the maximum likelihood estimate of the covariance matrix of normally distributed random vectors is obtained by matrix differentiation results for general matrices, not restricting the covariance matrix to being symmetric positive definite (SPD) or even just symmetric. A stationary point of the likelihood function is obtained—the stationary point is then observed to be SPD and it is concluded that this must be the unique solution to the maximization problem in the smaller domain of SPD matrices. Although the solution is correct its derivation is not, and some confusion may arise since the likelihood function attains arbitrarily large values when the covariance matrix is not restricted to being SPD. Naturally the stationary point obtained for general matrices was in fact a saddle point, and some further insight about the situation is provided by the monotonicity result presented for determinants here.

The maximum likelihood estimate

The likelihood function of N independent normally distributed random vectors \mathbf{e} with n real components is given by

$$L = (2\pi)^{-Nn/2} |\mathbf{R}|^{-N/2} \exp \left\{ -\frac{1}{2} \text{tr} (\mathbf{E}^T \mathbf{E} \mathbf{R}^{-1}) \right\} \quad (1)$$

where \mathbf{R} is the unknown covariance matrix. The matrix \mathbf{E} is formed from the observations according to

$$\mathbf{E}^T = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N]. \quad (2)$$

The standard way of obtaining the maximum likelihood estimates is by differentiating the logarithm of the likelihood function with respect to the estimated parameters using matrix differentiation rules for general matrices. This gives [see Goodwin and Payne (1977), p. 48, eqn 3.3.10]

$$\frac{\partial \ln L}{\partial \mathbf{R}} = -\frac{N}{2} \mathbf{R}^{-1} + \frac{1}{2} \mathbf{R}^{-1} \mathbf{E}^T \mathbf{E} \mathbf{R}^{-1} \quad (3)$$

* Received 19 December 1989; received in final form 12 June 1990. Recommended for publication in revised form by Editor W. S. Levine.

† Department of Chemical Engineering, Åbo Akademi, Biskopsgatan 8, SF-20500 Åbo, Finland.

‡ Author to whom all correspondence should be addressed.

The maximum likelihood estimate of \mathbf{R} is given by the root

$$\hat{\mathbf{R}} = \frac{1}{N} \mathbf{E}^T \mathbf{E}. \quad (4)$$

The saddle point nature of this symmetric root is seen as follows. We perturb the inverse solution with a real nonzero skew-symmetric matrix $\mathbf{H} = -\mathbf{H}^T$.

$$\mathbf{R}^{-1} = N(\mathbf{E}^T \mathbf{E})^{-1} + \Delta \mathbf{H} \quad (5)$$

where Δ is a real scalar. Observe that the trace within the exponential in equation (1) is unchanged by this perturbation, since the trace of a sum is the sum of traces, and the skew-symmetric real matrix $\mathbf{E} \mathbf{H} \mathbf{E}^T$ has zero diagonal elements whereby $\text{tr}(\mathbf{E}^T \mathbf{E} \mathbf{H}) = \text{tr}(\mathbf{E} \mathbf{H} \mathbf{E}^T) = 0$. Therefore

$$L(\mathbf{R}^{-1}) = L(\hat{\mathbf{R}}^{-1}) \left(\frac{|\hat{\mathbf{R}}^{-1} + \Delta \mathbf{H}|}{|\hat{\mathbf{R}}^{-1}|} \right)^{N/2} \quad (6)$$

Also the determinant inequality

$$|\hat{\mathbf{R}}^{-1} + \Delta \mathbf{H}| \geq |\hat{\mathbf{R}}^{-1}| \quad (7)$$

holds for all $\Delta \neq 0$ as $\hat{\mathbf{R}}^{-1}$ is SPD (see the Appendix) so that

$$L(\mathbf{R}^{-1}) \geq L(\hat{\mathbf{R}}^{-1}). \quad (8)$$

The likelihood function will thus be increased (monotonically with respect to $|\Delta|$) by perturbations about the stationary point with skew-symmetric matrices, and the stationary point given by equation (4) is only a saddle point (for general matrices as the domain).

A simple solution

The maximum likelihood estimate can be obtained rigorously by using differentiation rules for symmetric matrices (Graybill, 1983). However, the following elementary and concise approach is more appealing especially for classroom use.

Constrain \mathbf{R} to be SPD and assume $\mathbf{E}^T \mathbf{E}$ is invertible so that it is also SPD. Then, square roots of these matrices are defined (uniquely by requiring them to be SPD). Define the matrix

$$\mathbf{A} = (\mathbf{E}^T \mathbf{E})^{1/2} \mathbf{R}^{-1} (\mathbf{E}^T \mathbf{E})^{1/2} > 0 \quad (9)$$

and note that it has the same trace as $\mathbf{E}^T \mathbf{E} \mathbf{R}^{-1}$. The determinant of \mathbf{A} is related to that of \mathbf{R} by

$$|\mathbf{A}| = |\mathbf{E}^T \mathbf{E}| |\mathbf{R}^{-1}|. \quad (10)$$

Now maximization of equation (1) is equivalent to maximizing

$$f = |\mathbf{A}|^{N/2} \exp \left\{ -\frac{1}{2} \text{tr} (\mathbf{A}) \right\} \quad (11)$$

with respect to \mathbf{A} where \mathbf{A} is SPD.

Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the eigenvalues of \mathbf{A} —being SPD \mathbf{A}

can be diagonalized and all its eigenvalues are positive. Then

$$f = \left(\prod_{i=1}^n \lambda_i \right)^{N/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \lambda_i \right\} = \prod_{i=1}^n \lambda_i^{N/2} e^{-(\lambda_i/2)}. \tag{12}$$

(This equation would be equally valid for general symmetric matrices **A** (or **R**) and considering negative λ_i , clearly shows that no global maximum would ever be attained.) The stationary point is now obtained from the last expression by considering the factors separately:

$$\frac{d \lambda_i^{N/2} e^{-(\lambda_i/2)}}{d \lambda_i} = \lambda_i^{(N/2)-1} e^{-(\lambda_i/2)} \left(\frac{N}{2} - \lambda_i \right) = 0. \tag{13}$$

For the allowed eigenvalues $\lambda_i \in [0, \infty[$ the unique solution of equation (13) is

$$\lambda_i = N, \forall i. \tag{14}$$

Since the derivative of each factor changes sign just once, from positive to negative, the stationary point obtained is the global maximum (within the domain considered).

Now all the eigenvalues of **A** are equal. Note that the only matrix similar to a multiple of the identity is that multiple itself, and

$$\mathbf{A} = N\mathbf{I} = (\mathbf{E}^T \mathbf{E}) \mathbf{R} \tag{15}$$

The unique SPD maximum likelihood estimate of **R** is therefore

$$\hat{\mathbf{R}} = \frac{1}{N} \mathbf{E}^T \mathbf{E}. \tag{16}$$

Discussion

An elementary proof for the maximum likelihood estimate of the covariance matrix, for normally distributed random vectors, was presented. This proof, it is hoped, will supersede the less rigorous but more complicated proofs in some current textbooks. Aside from the main theme a determinant inequality, that the authors have not managed to find in the literature, is derived in the Appendix and used for illustrating the importance of proper consideration of the domain in optimization problems.

References

Goodwin, G. C. and R. L. Payne (1977). *Dynamic System Identification. Experimental Design and Data Analysis*. Academic Press, New York.
Graybill, F. A. (1983). *Matrices with Applications in Statistics*. Wadsworth, Belmont, CA.

Appendix

Let **S** be a real symmetric positive definite matrix and **H** be some nonzero real skew-symmetric matrix (**H** = -**H**^T). Here we show that

$$|\mathbf{S} + \Delta \mathbf{H}| > |\mathbf{S}| \tag{A.1}$$

for all values of the real scalar $\Delta \neq 0$, and in fact monotonically increases with the absolute magnitude of this perturbation parameter. (The reader may observe that the same proof is valid for the skew-Hermitian perturbation of a Hermitian matrix in the complex case, provided that absolute values of the determinants are taken).

Observe that

$$\mathbf{S} + \Delta \mathbf{H} = \mathbf{S}^{1/2} (\mathbf{I} + \Delta \mathbf{S}^{-1/2} \mathbf{H} \mathbf{S}^{-1/2}) \mathbf{S}^{1/2} \tag{A.2}$$

and by the product rule for determinants

$$|\mathbf{S} + \Delta \mathbf{H}| = |\mathbf{S}| \cdot |\mathbf{I} + \Delta \mathbf{G}| \tag{A.3}$$

with

$$\mathbf{G} = \mathbf{S}^{-1/2} \mathbf{H} \mathbf{S}^{-1/2} \tag{A.4}$$

Since **G** is skew-symmetric its eigenvalues are purely imaginary, and these are shifted by unity when the identity matrix is added:

$$|\mathbf{S} + \Delta \mathbf{H}| = |\mathbf{S}| \prod_{j=1}^n (1 + i \Delta \lambda_j), \tag{A.5}$$

where $i \lambda_j, j = 1, \dots, n$ are the eigenvalues of **G** and $i = \sqrt{-1}$. The product on the RHS is pure real since the LHS is, so taking the absolute value will at most change the sign. Shifting the absolute value to the factors of the product gives

$$|\mathbf{S}| \prod_{j=1}^n \sqrt{1 + \Delta^2 \lambda_j^2}, \tag{A.6}$$

which obviously is monotonically increasing with $|\Delta|$, strictly so since at least one of the eigenvalues is nonzero. Due to continuity with respect to Δ the RHS of (A.5) cannot jump to negative values as Δ moves away from zero; thus it stays positive and taking the absolute value does not even change sign. This proves that the LHS of (A.5) is also monotonically increasing with respect to the absolute value of Δ . The weaker result

$$|\mathbf{S} + \Delta \mathbf{H}| > |\mathbf{S}| \tag{A.7}$$

for all real $\Delta \neq 0$, follows from this strict monotonicity.

Q.E.D.

Real-time Computer Control: An Introduction*

Stuart Bennett

Reviewer: L. MOTUS

Institute of Cybernetics, Computer Division, Akadeemia tee
21, Tallinn 200108, Estonia.

THIS BOOK is much better than I expected it to be—I was sure that it was another book on control algorithms, emphasizing now and then the necessity of very rapid computations. Fortunately enough, this was not the case.

I think that the book presents a nice collection of knowledge areas, necessary to specify, design, implement, debug and test a real-time computer control system. The depth of presentation is suitable for the audience intended for the book—the final year undergraduate students and practising engineers. Still, I tend to think that an engineer can learn more from this book than a student: implicit connections between different knowledge areas may be a little bit difficult to discover without a practical control system background.

In the following, the book is reviewed chapter by chapter and subjective merits and demerits are listed for each chapter.

Chapter 1, Introduction to Real-time Systems, is perfect, especially the interpretation of real-time—in most cases real-time means that two or more systems (one being a computer-based system) with different inner time concepts are forced to cooperate by providing a common understanding of time. It was so refreshing to see that the widespread misconception of real-time as being almost equivalent to super-performance is no longer taught to students.

Chapter 2, Concepts of Computer Control, seems to be not quite balanced. The attention has been biased towards explaining different control strategies (sequence control, direct digital control, supervisory control), whereas the problems of control system as a whole (centralized computer control, hierarchical control system, distributed control system) are discussed superficially. In addition, I am not quite happy with the notion of distributed system—in the book it is interpreted more like distributed computing. I would personally use the term “distributed control system” as a common name for a system that may contain both hierarchical control and distributed computing elements. This chapter would have been a suitable place for explicitly stating how the following chapters are related to each other in the process of building a control system. However, there is no such statement in the book.

Chapter 3, Computer Hardware Requirements for Real-time Applications, is a very good explanation of how computer hardware functions (including also process interface and data communication in distributed system). The content of this chapter is closer to the description of the state-of-the-art of computer hardware than to the requirements. I would have expected more thorough discussion of standards, 24 lines is certainly not sufficient.

Chapter 4, DDC Control Algorithms and their Implementation, is another excellent part of this book. The reader is systematically introduced to the differences of coding a data processing algorithm and programming a real-time system. Slightly out of line of this book seem to be the details of DDC algorithms, e.g. improved forms for integral and

derivative calculations, z-transform. Example 4.5 contains an error (nextSampleInterval pro nextSampleTime).

Chapter 5, Design of Real-time System, gives a reasonable survey of the hot topics of software engineering environments for computer control systems. I cannot fully agree with separating hardware and software designs before requirements analysis; perhaps it would be better to prove consistency and noncontradiction of the system definition (requirements) first. In many cases it would be difficult to fix a feasible hardware configuration before we have complete requirements on the system behaviour. Also, in control systems it may often happen that it is necessary to implement some part of software by using dedicated hardware and again, this kind of a decision depends on the control algorithms and on the plant dynamics. I doubt that the system definition (see Fig. 5.1) and the indicated interactions (see Fig. 5.1 and Fig. 5.2) are sufficient for that. At least these problems should have been mentioned in this chapter.

Another problem worth mentioning in this chapter is connected with the multiple use of the system/software/hardware abstract model. Traditionally the abstract model is used for providing a better survey of the system/software/hardware to the designer in a hope that he/she can find discrepancies or errors in the description. Today the use of formal methods for discovering errors gains more and more popularity. Application of formal methods to system descriptions, in their turn, imposes new requirements to abstract models. This problem is not discussed in the book.

Chapter 6, Operating Systems, gives a good explanation of the basic parts of these systems. This explanation is necessary and sufficient for understanding the functioning of a real-time system. A large part of the information obtained from the design description (see the previous chapter) is often used to determine or modify some of the operating system properties (e.g. scheduling, queue lengths, etc). The summary of this chapter just mentions the problem. I think students would need a more detailed discussion of this subject.

Chapter 7, Concurrent programming, gives a pragmatic overview of parallel execution of programs—described are methods and primitives necessary for implementing multi-tasking (mutual exclusion and intertask communication problems). Presentation of methods and primitives is clear and understandable and is supplied with a number of examples. The given amount of information is sufficient for building not too demanding control systems. However, some of the practising engineers may have encountered more sophisticated concurrent programming problems, e.g. timing problems in intertask communication, or really concurrent implementation of some of the algorithms. For those it would be useful to know that concurrent programming means much more than is presented in this chapter.

The heading of Chapter 8, Real-time Languages, is quite a confusing one. Not much is said about real-time languages in it, in fact, the next chapter handles them more thoroughly. This chapter is about requirements to languages for programming large systems.

Chapter 9, Programming Languages, surveys some of the languages used for implementing computer control systems (Basic, Fortran, Pascal, Coral 66, RTL/2, Modula-2, Ada). I would have added PEARL to the list of surveyed languages. PEARL is used on the European continent and has a nice design supporting system (EPOS)—it provides a good example of systematic approach to the development of real-time systems.

On the whole the book presents a well-balanced collection of topics important for designing and implementing computer

* *Real-time Computer Control: An Introduction* by Stuart Bennett, in *Series in Systems and Control Engineering* (Series editor: M. J. Grimble), Prentice Hall International, Hemel Hempstead, U.K. (1988). ISBN 0-13-762485-9, 362 pp., \$82.95.

control systems. It is readable and the presentation is clear. The basic drawback of the book is that it does not merge the different areas of human knowledge into one; this is left for the reader. I am afraid that this task is not so easy for an undergraduate student.

In spite of this drawback the book is one of the best textbooks introducing real-time computer control that I have read.

About the reviewer

Leo Motus has Ph.D. equivalents in stochastic control and software engineering. He has been Chairman of the IFAC working group on Distributed Computer Control Systems since 1987, is a Vice-chairman of IFAC TC on Computers, and is a candidate for the post of TC Computer Chairman for the period 1990–1993.

Computer Control of Machines and Processes*

John G. Bollinger and Neil A. Duffei

Reviewer: C. SCHMID

Ruhr-Universität Bochum, Lehrstuhl für Elektrische Steuerung und Regelung, Postfach 102148, D-4630 Bochum 1, Federal Republic of Germany.

THIS BOOK is an excellent one for undergraduate and graduate students of all engineering disciplines interested in computer control, as well as being a superb reference book for industrial people attending refresher courses on this subject. One of its striking features is that little prior knowledge is demanded from its reader/user. Its spectrum of topics, which are of special importance in computer control, is indeed exceedingly broad. The book has naturally evolved from courses presented at the University of Wisconsin (Madison) by the authors since the mid-1960s.

In its more than 600 pages, the book covers all the relevant aspects of computer control. The twelve chapters essentially give a nonmathematical introduction. After the brief introductory Chapter 1, which sets the rest of the book in proper perspective by dealing with the history of computer control and the explosion in the application of electronic technology since the 1960s, it goes on to present chapters on elements of discrete-time modelling, system response generation, discrete-controller design, control computer hardware and software, computer interfacing, sensors, command generation, sequential logical control, process modelling, analysis and design. Finally, the book closes with an appendix on the state-variable approach.

Chapter 2 starts with a discussion on basic process types and develops the idea of representing components of closed-loop systems using difference equations and discrete-time transfer functions, without applying any transformation techniques (which will be first introduced towards the end of the book in Chapter 11). After reading this chapter, one is on the level to understand system input-output behaviour and to analyse the stability of a system. However, the discussion of stability is different from the standards of control theory! As the stability analysis is based on the roots of the characteristic equation in terms of the backward-shift operator, all discussions about stability are treated in a domain mirrored at the unit circle. This may cause the novice reader to be confused when reading other books recommended in the bibliography attached to this chapter. Two fundamental concepts are introduced in Chapter 3: The sampling of signals generating time series, and the use of the time-shift operator to generate and determine system responses. The first design of a simple controller can be found in Chapter 4, which focuses on the desired closed-loop responses to specific inputs. The discussion about the selection of the sample period and the design of feedforward, cascade and noninteracting control in interactive plants

completes the first part of this book. These four chapters give an excellent introduction to discrete-time control.

The second part of this book deals with the computer. Chapter 5 is dedicated to computer hardware and software. It presents important aspects of computer architecture and operation in a way that is independent of computer manufacturer or model. This includes binary logic, basic computer hardware, the concepts of instructions and data, input/output, interrupts and programming at the assembly language level. A set of high-level language procedures is defined such that the principle of closed-loop computer control can be shown at the end of the chapter. This chapter concludes with a discussion of how closed-loop control functions can be organized on a control computer. The material and examples are well elaborated and their sequence reflects the various levels of abstraction in which designers of computer-control systems must carry out their work. Linking to external devices is the main topic of Chapter 6. The range is from analogue conversion to address decoding, device selection and interrupt interfacing. The short discussion here draws on a simplified computer architecture and is sufficient for understanding. Chapter 7 describes a spectrum of sensors that are often found in computer control systems for machines and processes. Chapter 8 adds some nonessential aspects of command signal generation in control. The implementation of logic control and the solution of Boolean equations are discussed in Chapter 9. Ladder diagrams, which are the means to describe the solution of logic on a computer, are treated. In addition, a number of design methods for logic control are described—including the use of flowcharts, switching tables and state diagrams. This chapter gives only a rough sketch about the principles of sequential control and the use of programmable logic controllers. Chapters 1–9 can be covered in one semester at an introductory level.

The third part of the book portrays more control techniques, the crucial point being the application of more sophisticated system techniques to computer control. Chapter 10 reviews a number of approaches to process modelling—from physical modelling to a mathematical one. Step-response and least-squares techniques of process model identification are used. Through Chapter 11, where he can find the bases of transformation into the frequency domain, the reader will be able to gather much more insight into dynamics. The concepts of ideal samplers, hold elements and the representation of a sample sequence are also introduced here. After the analysis section in Chapter 12, the design of controllers in the frequency domain is discussed both for continuous- and discrete-time systems, illustrating their similarities and differences. An appendix gives a brief review of state-variable methods for the analysis and design, including state estimators for continuous- and discrete-time systems.

Throughout the textbook, technology-based information has been strictly avoided. The simplified computer hardware which is used allows fundamental concepts to be illustrated,

* *Computer Control of Machines and Processes* by John G. Bollinger and Neil A. Duffei. Addison-Wesley, Reading, MA (1988). ISBN 0-201-10645-0, \$53.75.

and also allows system-level issues to be discussed without describing too complex computer architectures. The authors include PASCAL programs at many locations in the text to illustrate the use of computers. The programs are greatly simplified to improve the illustration of important concepts. Various manufacturing-related examples of processes are used to illustrate the analysis, design and implementation techniques that have been introduced. Most chapters start and/or close with an example, the layout being uniform throughout the book. Also, each chapter starts with an introduction which contains a precise description of the goals. At the end, a more complex, but still illustrative, example comprises the essentials of the chapter which are additionally highlighted in a special summary section. A bibliography is attached to each chapter, and most of its entries are books which, although not referred to directly in the text, may be used to supplement the book if additional information is required. In order to become more engaged, a well-elaborated list of many further interesting examples can be found at the end of each chapter in form of exercises.

The printing is of high quality, and the figures are

enlightening where they should explain and illustrative where they have to describe a real object. The whole work shows the great teaching experience of the authors on this subject—towards which of course, students' feedback enjoys great credit.

The book is highly readable and well organized. It may be concluded that this textbook is intended where there are great differences in the academic background of students who enrol in computer control courses.

About the reviewer

Dr C. Schmid received the Dipl.-Ing. degree in mechanical engineering from the University of Stuttgart in 1972 and Dr.-Ing. degree in electrical engineering from the Ruhr-University Bochum in 1979, where he has been a lecturer since 1980. His main teaching activities are simulation techniques and computer-aided control system analysis and design, and his research interests are in adaptive control and the practical aspects of CAD of control systems. He is the author of a well-known industrial CAD system.

Industrial Control Electronics: Applications and Design*

J. Michael Jacob

Reviewer: H. RAKE

Institute of Automatic Control, Aachen University of Technology, Aachen, Germany.

As in many other disciplines of our world relying on science and technology, a noticeable gap has developed in control engineering between (control) theory and (electronic) practice. The present book, by J. M. Jacob, intends to bridge this gap. Written for electrical engineering technology undergraduates it seems to be well suited to introduce control engineers who are not electronics specialists to the innards of the electronic hardware which they will have to use in order to close control loops.

The book's 585 pages are divided into eight chapters and three appendices. About three quarters of the pages are devoted to material suggested by the book's title, and this material is presented in professional breadth and depth. The remaining quarter of the book is far less recommendable. It seems to have been written for the benefit of the electrical engineering technologist to provide some insight into the world of control engineering. From a control engineer's point of view, this cannot be rated as a success. (By the way, did any reader encounter a "Bode Plot" in theory or practice? The reviewer, remembering the American H. W. Bode, did not.)

If treated as a one-way bridge, namely from control engineering to electronics, the book deserves much praise. The fascinating subject of transducers is covered in a well-structured chapter giving valuable details about almost every type of some significance. The reviewer found only the coriolis-type mass-flow transducers missing, which seems to

be a minor shortcoming. Signal conditioning and transmission is treated including the peculiarities of converters, isolation circuits and cabling. Design principles of analog controllers are explained in such a way that the working of these devices including derivative overrun, integral windup and bumpless transfer from manual to automatic operation become transparent. The corresponding chapter on digital control gives very detailed and valuable information about A/D- and D/A-conversion while the rest of the chapter looks rather sketchy and a few pages devoted to programmable controllers could have been omitted. Power interfaces are the last electronic devices needed for closing the control loop. Switch-mode amplifiers and thyristor devices are described in useful detail, showing how to make low-power signals effective in a world of high voltages and high currents compelled to energy efficiency.

In concluding, the reviewer highly recommends the book to all those members of the control community who are not too familiar with electronics and want to know more about the how and why of electronic devices and systems. They will surely benefit from the strengths of this book and be barely irritated by its (perhaps inevitable) weaknesses.

About the reviewer

Heinrich Rake graduated in mechanical engineering from Hannover Technical University. He received his Ph.D. in 1965 for a dissertation on self-adaptive controllers. Working at the Institute of Automatic Control of Aachen Technical University since 1966, he was appointed head of this institute and Professor of Control Engineering in 1977. Main areas of research are model-based adaptive switching control, control of air conditioning plants and industrial robots, modelling of discrete-event systems, identification and parameter estimation methods and application of automatic control principles to a variety of industrial processes.

* *Industrial Control Electronics: Applications and Design* by J. Michael Jacob. Prentice-Hall International, Hemel Hempstead, U.K. (1988). ISBN 13-459322-7, \$31.95.

Power Hydraulics*

Michael J. Pinches and John G. Ashby

Reviewer: H. RAMON

Automatic Control Laboratory, State University of Ghent,
Grotesteenweg Noord 2, B-9710 Gent Zwijnaarde, Belgium.

ALTHOUGH PRIMITIVE hydraulic power (control) goes back to ancient times, it was the theoretical work on hydrostatics of the Frenchman Pascal, published in 1648, that opened new perspectives in power hydraulics. The first practical realization of Pascal's Law was devised more than a century later by Bramah, who demonstrated the working of a simple hydraulic press. From that moment on, fluid power applications were growing constantly, with a temporarily diminishing growth rate at the end of the nineteenth century due to the rise of electrical power transmission.

However, only since World War II, with the introduction of servovalve technology, have modern developments in hydraulic power control expanded enormously. Starting from early applications in airplanes, fighters and missiles, it spread over almost the whole field of mechanical engineering and mechatronics: from mining, shipbuilding and the steel industry to automotive engineering, robotics, aeronautics... This expansion can mainly be explained by the unique features of hydraulic power, such as: the high safety and simple maintenance of the hydraulic system; good dynamic characteristics with fast starts, stops and speed reversals; high stiffness properties of hydraulic actuators which give little drop in speed with increased load; and the versatility of its usage, e.g. in compact and light devices that can develop hydraulic power varying from a few kilowatts to several megawatts; hydraulic fluids which act as lubricants and carry away superfluous heat, generated during action, to a heat exchanger (both of these properties resulting in a low wear of components).

In many industrial applications, the positive features of hydraulic systems greatly surpass their disadvantages, which include: low efficiency, the high cost of hydraulic parts and their high sensitivity to dirt and contamination in the fluid which gives rise to damage or silting of expensive components.

During the last three decades, fluid power hydraulics evolved from a more empirical science, with emphasis on fluid mechanics, to a highly technological and interdisciplinary subject in applied science. It asks from the designer, besides an intimate insight into component design and component interaction, a good working knowledge of engineering mechanics, hydraulics, electricity and electronics, classical and modern control theory, instrumentation and computer science.

It is clear that a "complete book" about all aspects of hydraulic power systems would result in an issue of thousands of pages. Therefore authors have to make a logical selection between some topics of these disciplines and join them together in a coherent and readable work. The choice lies between two extremes, and depends on the target group.

The first extreme includes all books which keep mathematics and calculations as simple as possible, but try to give a clear physical insight into the operation and interaction of all accessories in hydraulic circuits (pumps, valves, actuators, filters, conduits, accumulators, hydraulic fluids etc.). To be complete in their domain, they should also describe rules and practical hints for hydraulic system design and maintenance. Consequently it should make them not only useful as an introductory course for junior students, but also of help to practising engineers and even to craftsmen.

In the second extreme, all recent and advanced topics in hydraulic power system design should get their turn. This certainly starts with a mathematical representation of the dynamic characteristics of different lumped and distributed system components such as valves, pumps, motors, actuators and transmission lines. Since most of them show a highly nonlinear behaviour, a nonlinear analysis seems inevitable beside a linearized description. These models are not only useful for accurate simulation purposes, which can be employed in design studies, but also to predict and detect the undesirable behaviour of some devices under certain working conditions.

The next important topic concerns the use of modern system theory in power hydraulics. Due to their excellent features, hydraulic devices are utilized more and more in complex mechanical control systems, e.g. in vibration control on elastic structures, extremely precise position, speed- and force control of flexible robots, active and semi-active suspensions in automobiles; these ask for advanced control algorithms. For this reason, state-space and multivariable frequency domain techniques, nonlinear-, digital- and robust control, adaptive- and self-tuning control, model reference control and (on-line) system identification, are becoming indispensable methods in hydraulic power control.

It is evident that the ability to apply these techniques together with their advantages, restrictions, and shortcomings should be described rigorously in close relation with hydraulic system characteristics. As a consequence this kind of book would be of interest to (under)graduate students, interested researchers and practising engineers, who need a profound knowledge of the subject.

Turning to the existing literature in this field, one notices that there mainly exist two categories of book. In the first, the material is handled in a descriptive way (Banks and Banks, 1988; Rexroth, 1981, 1986, 1988, 1989a, b) while in the second, most authors treat the subject at an intermediate level. This means that they try to give a well balanced compromise between a descriptive and a quantitative approach in which design calculations are normally based on a linear analysis and classical control theory (Merritt, 1967; Stringer, 1976). Sometimes, more advanced topics are discussed in a small part of the book (Anderson, 1988; McCloy and Martin, 1980). However, a consistent and rigorous treatment of advanced theory through the whole text is rather exceptional (Watton, 1989) and most of the time the specialized reader has to be satisfied with articles and PhD dissertations (Dietz, 1988; Faulhaber, 1985; Feuser, 1983; Quetting, 1982).

The present book, *Power Hydraulics* by Michael J. Pinches and John G. Ashby may be considered as an excellent contribution to fluid power hydraulics. In 400 pages it demonstrates, from a practical viewpoint, the simple calculations, circuitry and component selection involved in system design.

Chapter I starts with a very short introduction to hydraulic principles, followed by an explanation of the most important hydraulic symbols. Chapters II, III, IV and V discuss the properties, operation and construction of hydraulic equipment (pumps, hydraulic valves, actuators, filters, hydraulic fluids, tubes and accessories). Chapter VI establishes design criteria based on elementary formulae and rules. Reservoirs and accumulators are treated here, and two detailed design studies are presented. Chapter VII handles in depth several procedures and actions which benefit good maintenance, together with a guidance on trouble shooting, illustrated with six examples. In Chapter VIII, linear models of valve and pump servo systems are derived and their response to step, ramp and sinusoidal inputs are calculated. At the end proportional valve technology is stated and compared with

* *Power Hydraulics* by Michael J. Pinches and John G. Ashby. Prentice-Hall International, Hemel Hempstead, U.K. (1988). ISBN 13-687443-6. \$76.00, £39.95.

servovalves. Finally, there is an appendix with exercises and solutions.

In the first seven chapters the authors describe in a well structured and almost encyclopaedic way different types of components with their principles of operation, repeatedly illustrated with figures and hydraulic circuits. The number of utilized formulae is kept to an absolute minimum and simple algebraic computations are restricted to the many interesting examples which clarify design rules in the text.

No previous knowledge of hydraulics or calculus is needed, except in the last chapter which presumes an elementary knowledge of ordinary differential equations and Laplace transforms. Each chapter is interlaced with numerous tables, diagrams and hundreds of practical hints which make the book invaluable for practising people in the field. As a consequence, the aim of the authors to cater to a broad public of equipment purchasers, craftsmen, practising engineers, lecturers and students, is certainly reached. The book is also carefully edited, only SI-units are used, and the choice between a hard cover or a low cost student edition is possible.

Since nothing is perfect, it is normal that the book displays some shortcomings. The weakest link is certainly the chapter about pumps. For a few types of pump, the explanations with corresponding figures are not clear enough to understand well their working principles (the internal gear pump of p. 19 and the gerotor pump of p. 20). The drawing of a radial piston pump is overexaggeratedly detailed, which is not relevant to the text, and Fig. 2.6 of an axial piston pump seems to be incorrect.

I want to finish with a strictly personal remark: it is a pity that the authors did not speak of the so important valve coefficients (flow gain, flow-pressure coefficient, pressure sensitivity), pressure sensitivity curves and dynamic characteristics of servovalves with the aid of Bode diagrams. It would fit perfectly in chapter 8 and would make the book even more valuable as an "introductory" university course.

References

- Anderson, W. R. (1988). *Controlling Electrohydraulic Systems*. Marcel Dekker, New York, 1988.
 Banks, D. D. and D. S. Banks (1988). *Industrial Hydraulic Systems: An Introduction*. Prentice-Hall, New York.
 Dietz, U. (1988). *Nichtlineare Zustandsregler für Elektro-Hydraulische Servoantriebe*, Fortschritt-Berichte VDI, Reihe 8, Nr. 155. VDI-Verlag, Düsseldorf.
 Faulhaber, S. (1985). *Lageregelungen für Hydraulische*

Servoantriebe, Fortschritt-Berichte VDI, Reihe 8, Nr. 84. VDI-Verlag, Düsseldorf.

Feuser, A. (1983). Ein Beitrag zur Auslegung Ventilesteuerten Hydraulischer Vorschubantriebe im Lageregelkreis. Doktor-Ingenieur Dissertation, Technischen Fakultät der Universität Erlangen-Nürnberg.

McCloy, D. and H. R. Martin (1980). *Control of Fluid Power; Analysis and Design*, 2nd ed Ellis Horwood, Chichester, U.K.

Merritt, H. E. (1967). *Hydraulic Control Systems*. Wiley, New York.

Quetting, P. (1982). *Zustandsregelung eines nichtlinearen Systems am Beispiel eines elektro-hydraulischen Stellantriebs*, Fortschritt-Berichte VDI, Reihe 8, Nr. 48. VDI-Verlag, Düsseldorf.

Retroth (Mannesmann) (1981). *Der Hydraulik Trainer*, Band 1: Grundlagen, Lohr am Main, Band 2, *Proportional- und Servoventil-Technik* Lohr am Main, 1986; Band 3, *Projektierte und Konstruktion von Hydraulanlagen*, Lohr am Main, 1988; Band 4, *Technik der 2-Wege-Einbauventile*, Lohr am Main, 1989a; Band 5, *Fluidtechnik von A bis Z*, Lohr am Main, 1989b.

Stringer, J. D. (1976). *Hydraulic Systems Analysis*. Macmillan, London.

Watton, J. (1989). *Fluid Power Systems, Modeling, Simulation, Analog and Microcomputer Control*. Prentice-Hall International, Hemel Hempstead, U.K.

About the reviewer

Herman Ramon worked as an agricultural engineer, specializing in mechanics, at the State University of Ghent, Belgium during 1982. From 1983-1985, he was an assistant at the Free University of Brussels Department of Hydrology, where his main task was teaching exercises in subsurface hydrology. He then worked at a private firm as project manager in the feedmill industry, being responsible for the development of mathematical models and the application of (non-)linear optimization techniques. Since 1988, Mr Ramon has been a research engineer, partly at the State University of Ghent, Laboratory of Automatic Control, and partly at the Catholic University of Louvain, Department of Agricultural Engineering, while preparing a PhD thesis on the controlling of undesirable flexible- and rigid-body motions in agricultural machinery with electro-hydraulic actuators.

Random Signals and Systems*

Richard E. Mortensen

Reviewer: J. F. BARRETT

Department of Mathematics and Statistics, Paisley College of Technology, Paisley, Renfrewshire PA12BE, Scotland, U.K.

THE PURPOSE of the book as stated in the preface is to serve as a textbook at the University of California at Los Angeles, for either a senior level undergraduate introductory course in stochastic processes or for a first year graduate level follow-up course. These courses are prerequisites for graduate courses in control and communication systems engineering.

For the requirement of the undergraduate course, the book aims to bring computer-oriented students, familiar

mainly with discrete mathematics, to grips with the more abstract style of mathematics used in current electrical engineering research. Such students will follow chapters 2, 4, 5, 6 (i.e. Gaussian distributions, discrete random sequences, Gaussian processes, filtering, power spectra) with a selection from other chapters (e.g. multidimensional Gaussian distributions, Volterra series, Markov processes).

For the requirement of the graduate course there is, in addition to this, material on applications of Hilbert space theory to random variables and the Karhunen-Loève expansion (Helstrom, 1960), which serves as a preparation for a more thorough course in measure theory and functional analysis. Supplementary topics such as estimation theory and state-space theory are available.

The book is planned to include a range of topics for flexibility (a "smorgasbord" as the author puts it) with illustrations taken in the main from radar and communication theory. It is written in an informal and discursive style to help maintain the students' attention. Exercises are available

* *Random Signals and Systems* by Richard E. Mortensen. Wiley, Chichester, (1987); hardback and paperback editions available. ISBN 4718 43 644, £31.40.

at the end of the chapters though without answers. A detailed analysis of the contents follows.

In Chapter 1, Discussion of Probability and Stochastic Processes, the basic concepts of probability are assumed to be partly familiar to the student from a previous course. According to the preface this chapter does not form part of the essential course material and so presumably is background material to be used at the lecturer's discretion.

The normal abstract definition of probability is given as a trio (Ω, \mathcal{A}, P) consisting of a sample space Ω , an algebra of admissible subsets \mathcal{A} and a probability measure P over \mathcal{A} . A slight notational ambiguity is present in that no distinction is made between a set and the event that a random variable takes a value in that set. The following sections introduce the basic definitions of probability theory in a concise and clear way. An interesting feature is the introduction of the Hilbert space of random variables which gives a geometrical insight into the abstract concepts.

The final section gives the basic definitions for a random process, a Markov process, and a Gaussian process. This is rather jumping ahead and these definitions would fit more naturally into later chapters. The definition of a random process as a family $\{X_t; t \in T\}$ of random variables all defined on the same probability trio (Ω, \mathcal{A}, P) is inadequate (strictly speaking of course), the probability distribution being required over the whole product space.

In Chapter 2, The Gaussian Distribution in One and Two Dimensions, after recalling the main facts of the one dimension distribution, the two dimensional density function is quoted in terms of its covariance matrix c_{ij} and its inverse s_{ij} . The remainder of this short chapter works out, in a numerical case, the integral of the two-dimensional density function over the positive quadrant, the object of the exercise being to provide practice in important elementary techniques such as completing the square and the use of polar coordinates.

Much basic material has been omitted. The student has presumably met simple least squares and regression previously. Should this not now be put into context with the two-dimensional Gaussian distribution which would also clarify the role of the conditional probability density function and prepare the ground for the more complicated multidimensional case treated in the next chapter? The Pearson normalized correlation coefficient ρ in terms of which the two-dimensional density is most frequently expressed is nowhere mentioned. Surely also it is worthwhile to calculate the characteristic function (for both one- and two-dimensional cases) and make use of it to find moments. This is in fact done in a piecemeal way throughout the book in the exercises. Another relevant topic would be the symmetric two-dimensional distribution and the associated Rayleigh distribution—very important in communications and also used in the FORTRAN program of Appendix 2.

Chapter 3 is titled The Multidimensional Gaussian Distribution. It deals with the multidimensional density function, the bivariate case, and the Bayesian estimation theory, and provides optional extra material.

The multidimensional density having been defined, it is shown how the Gaussian quadratic form may be reduced to diagonal form using the transformation of Appendix 1. The chapter continues with its main idea which is the derivation of the conditional density function $f(x/y)$ for a bivariate Gaussian vector (x, y) . To this end there is given a detailed derivation of the inverse of the bivariate covariance matrix using a matrix extension of the Gauss-Jordan method, well-known in elementary numerical analysis.

This is certainly an interesting, easily understandable, and stimulating exercise extending the scope of known methods. It is however quite lengthy and should perhaps give way to more essential material. A byproduct of this calculation is the Matrix Inversion Lemma, well-known in Kalman theory. The calculation is then used to complete the computation of the conditional density $f(x/y)$ and so finally to find conditional mean and variance to be used for Bayesian estimation.

Bayesian estimation is treated in the next section. The use of the word "Bayesian" is nonstandard, an estimate

$\hat{X} = g(Y)$ of one random variable in terms of another being called Bayesian in view of the fact that X and Y are taken to be both random with a joint density function $f(X, Y)$. Surely the point of the Bayesian view is that X is estimated, not by a single value \hat{X} but by a probability distribution? The single value estimate occurs in the point estimation theory of R. A. Fisher (1925) proposed as an alternative to the Bayesian theory. The book is therefore using a hybrid version of these two theories.

The chapter continues with determination of the optimum estimator to minimize a loss function $L(\hat{X} - X)$ of the estimation error $\hat{X} - X$, the most important case being that when the loss function is the Euclidean norm giving the Minimum Mean Squared Error Criterion. The known result is then proved that the optimum MMSE value of X given Y is the conditional expectation of X given Y . The proof is by a nonlinear extension of the completing-the-square technique which is more incisive than the common calculus of variations method.

The final part of the chapter deals, very briefly indeed, with the all important Gaussian case where we find the standard multidimensional regression formulae (although they are not called by this name). This approach bypasses the conventional least-squares theory which is unfortunate in view of its importance and the fact that it ties up so well with the Hilbert space emphasis of the book.

Chapter 4 discusses Finite Random Sequences. According to the preface, it forms one of the core chapters. It aims to introduce computer-oriented students to the concept of a finite Gaussian sequence in a way suitable for the computer generation of such sequences. The two main sections have the titles "The Successive Viewpoint" and "The Simultaneous Viewpoint". The first has in mind the successive generation of values of independent Gaussian random variables and the formation of their linear combinations, i.e. it leads towards the concept of filtered Gaussian white noise, which is then discussed, its values being thought of as being generated by the algorithm given in Appendix 2 (which is actually of course, pseudo-random). The notion of an ensemble of possible realizations is considered and finally it is shown how a Gaussian sequence having a given covariance matrix C may be generated by factorizing C in the form LDL^T as in Chapter 2 using the values of L , which is lower triangular, for multipliers of the white noise for causal generation of the required sequence. The following section introduces the contrasting "simultaneous viewpoint" in which the sequence is viewed as a whole and represented by a random vector having a given mean and covariance matrix. In this way the student has been prepared for the concept of the infinite random vector i.e. stochastic process, discussed in the next chapter.

This is a novel approach to the teaching of random processes suitable for a mathematical laboratory class. It is of course limited to Gaussian sequences and it should be emphasized that other processes exist. The Gauss-Markov process and autoregressive scheme would have fitted rather naturally into this chapter; they are mentioned later. The insertion of a few graphs would aid the imagination.

Chapter 5, Stationary Random Sequences, deals with discrete-time random processes, in particular linearly filtered white noise and its second-order properties of covariance and spectrum.

The previous chapter has prepared the way for the definition of a random process as an infinite vector $\{X_k\}$. There follow the usual definitions of mean and covariance. For a stationary process the covariance becomes the autocovariance function of the time difference. An extended example discusses the properties of filtered white noise including calculation of output autocovariance and, for more than one process, cross-covariance. Power spectral density is defined by the direct method as the Fourier transform of the autocovariance function, its positivity being demonstrated by a quotation of Bochner's (1932) theorem. After a digression on aspects of linear system theory, input-output relations for spectral densities are derived and finally the spectral factorization of rational spectral densities is discussed.

All this calls for little comment except possibly for a

remark on the definition of spectral density. The definition by Fourier transformation of autocovariance function does not make clear either the phenomenon of aliasing or the significance of spectral density for the analysis of signal records. Since the book intends, according to the preface, to provide basic knowledge for signal processing, some explanation of these very important ideas would be in order.

Chapter 6—Continuous-Time Stationary Gaussian and Second Order Processes—extends the concepts of the last chapter to continuous-time systems. It makes sense educationally to treat the discrete case first as the definition of the continuous-time process can readily be accepted once the discrete-time process has been understood. These follow sections on the definition of second-order quantities including spectral density, the Laplace transform and linear systems, and input-output relations for linear systems leading to those for spectral densities. After a note on the Paley-Wiener criterion, there follow sections on ergodicity and the frequency interpretation of power for finite signals. All this is rather standard. The interesting topics for comment are spectral density and ergodicity.

Power spectral density is defined, as in the discrete-time case, directly as the Fourier transform of the autocovariance function, its positivity being demonstrated by quotation of Bochner's (1932) theorem. This is the method of definition of Khinchin (1931) and is usual in the mathematical-statistical literature. In electrical engineering the traditional definition is based on the frequency analysis of records of length T as $T \rightarrow \infty$. This is Carson's method used later by Rice, Middleton and closely related to Wiener's work. Although less tidy mathematically than the Khinchin method, it does have the advantage of relating spectral density to frequency analysis of data records. Getting back to the Carson definition from the Khinchin definition is not very difficult but is a "must" in view of the importance of the frequency analysis interpretation for signal analysis.

The section on ergodicity makes difficult reading. In view of the book's method of introducing random processes through computer algorithms a quite suggestive picture of the typical ergodic behaviour could no doubt have been given by considering the behaviour of simple number-theoretic algorithms of the type used to generate pseudo-random sequences. The traditional treatment of ergodicity in most textbooks is certainly not especially enlightening and, as a result, the topic of ergodicity usually has an air of mystery. The thoughts on this subject originate mainly from the Wiener-Lee school at M.I.T. Wiener himself was much concerned with the mathematics of ergodicity and his ideas on the subject (notoriously difficult to follow in the original!) have passed into folklore. The reviewer believes that the significance of his ideas is commonly misinterpreted. For it is implicit in Wiener's work, even if it is not stated in so many words, that the nonsingular stationary Gaussian process, i.e. filtered stationary white Gaussian noise, is in fact ergodic being derived from Brownian motion. This being so, the ergodic assumption is only necessary in other situations, e.g. when the Gaussian theory is applied to processes with assumed second-order stationarity. Since the theory deals so much with nonsingular Gaussian processes (as in this book) the ergodic assumption is usually unnecessary.

Chapter 7, Nonstationary Continuous-Time Processes, describes state-space representation and its extension to the nonstationary case, an extension which, as is known, goes through with ease.

Firstly, a linear system having rational transfer function $H(s)$ is represented by equations in state-space form whose solution is derived. The general time-varying state-space equations are then formulated and solved. The exposition is somewhat brief. Some of the basic facts about systems of linear differential equations would have been relevant leading to the semigroup property used in the derivation. Turning to the noise case there is given a detailed calculation of output covariance matrices for white noise input. The calculations cover $3\frac{1}{2}$ sides and although quite standard they would doubtless look formidable on first encounter by a student. They could well be simplified with quotation of

some of the equations since no application is made of them. In fact some illustration is needed to motivate the equations.

Chapter 8—Additional Topics in the Study of Continuous-Time Processes—is devoted to expansions of time functions on a finite time interval and their applications with emphasis on the use of Hilbert space ideas. The two types of expansion considered are those of Karhunen-Loève and Fourier. The chapter starts with a description of the Hilbert space of functions on a finite time interval with a concise explanation of the idea of a self-adjoint integral operator. These ideas are applied to the Karhunen-Loève expansion and illustrated by the case of Brownian motion which actually results in a K-L expansion which is a Fourier sine series. A more interesting case is provided by the Gauss-Markov process which is briefly introduced as an example at the end of the chapter. After some further discussion between the K-L expansion and the Fourier expansion, the theoretical ideas are illustrated by the important example of decoding a binary sequence in noise. A further illustration of the K-L expansion discussed is the work of Landau and Pollock (1961) relating to the time-frequency uncertainty principle where the kernel in the K-L integral equation is the sinc function.

The earlier part of the chapter reads very well, the section on the self-adjoint operator being especially well done. The two illustrations however need careful following, the first being decidedly technical and the second of necessity omitting all details. The use of communication problems to motivate and illustrate Hilbert space ideas is one of the appealing aspects of the book.

Chapter 9, Linear Systems in Conjunction with Memoryless Nonlinear Devices, deals with ideas originating in radar signal detection, the main part discussing the introduction of instantaneous nonlinear operations into linear systems theory, and the remaining part the radar uncertainty principle. The chapter starts with the description of the recovery of an amplitude modulated signal by a square-law detector followed by a linear low-pass filter. The theory is given for the deterministic case and then the stochastic case is considered. Here there is a curious error. It is stated that we run into trouble if, with $M(t)$ as a stationary message process, we try to take the spectrum of a process $(1 + M(t)) \cos \omega_c t$ because it is nonstationary. Now the same criticism would apply to a single sinusoid which certainly possesses a spectrum; the trouble is due not to nonstationarity but to nonergodicity. We must work with time averages—when this is done, the troublesome term $\cos \omega_c t \cos \omega_c s$ in equation (4) of page 147 becomes time-averaged to give $\frac{1}{2} \cos \omega_c (t - s)$. Then taking the spectrum we find the double humped form which is derived in the further development by the ad hoc assumption of page 148. The account given in pp. 150-152 could no doubt be simplified.

The difficulty of finding output probability density when nonlinear operations are present leads on to the topic of Wiener-Volterra series which is illustrated by the previously discussed square-law detector and then more generally by the model of a linear system followed by a memoryless nonlinearity, with expressions for the kernels in the case of a saturating nonlinearity. A somewhat more interesting illustration would have arisen in the standard representation for a receiver—linear system plus memoryless nonlinearity plus linear system.

The remainder of the chapter deals with ideas due to Gabor (1946) and others on complex signal representation and the radar uncertainty principle. The Hilbert transform theory used here is tricky even for the expert and in attempting generality the book fails to give a readable account. The suitability of this subject must be open to doubt. The actual derivation of the radar uncertainty principle is however quite straightforward given the possibility of complex presentation.

Chapter 10, Nonstationary Random Processes, deals with the Gauss-Markov process, generalizing it successively from scalar to matrix case and from stationary to nonstationary case. The scalar case with which the chapter starts could well have been discussed earlier in Chapter 4 where it logically

belongs. The interesting fact about it for the present chapter is the possibility of putting it in matrix form which leads on to the generalization to vector-valued sequences. This is used to solve the general recursive difference equation and to calculate means and covariances. The state-space equation with white noise forcing is finally discussed in its time-varying form. The chapter is quite straightforward: it is systematically developed and well written. It only lacks some interesting illustration of the theory.

In Chapter 11, Discrete-Time Kalman Filtering, the stated aim is to give an initial acquaintance with the Kalman filter without going into all details or its computer implementation, as well as to introduce some related concepts.

The problem is formulated in general terms for a scalar message process. The LDL^T factorization algorithm of Appendix 1 applied to the covariance matrix of the sequence of observations is used in its recursive form to calculate the innovation corresponding to a new observation. From this the updating half of the Kalman theory follows. The other half, the propagate formula, follows in the usual way from the Markov model. There follow some further comments on the LDL^T factorization in sequential form and on the Kalman filter equations, the standard formulae for the difference equations for the covariance matrices being derived.

The use of the LDL^T transformation makes this an interesting nontraditional approach though limited to a scalar message process. It might be expected to have computational advantages over the normal approach but these are not mentioned. Perhaps it would have helped to have clarified in detail the relation of the method to the traditional orthogonal projection method, especially in view of the Hilbert space interest of the book. As a first acquaintance with the Kalman filter this chapter would be tough going. The Kalman filter is not an easy thing to understand. In teaching it there is much to be said for following the traditional approach illustrated by simple examples and not avoiding numerical illustrative examples which can help to give an initial understanding.

Of the Appendices, the first proves that a positive definite symmetrical matrix C can be represented as LDL^T where L is a unique lower triangular matrix having 1 on the main diagonal. Moreover the factorization of C into this form can be performed sequentially. The result is a generalization of the known factorization into upper and lower triangular matrices. It is a useful result for systems analysis which does not appear to have been emphasized elsewhere.

The second appendix collects together a number of statistical ideas of use in the text such as a FORTRAN program for Gaussian random numbers, parameter estimation, statistics of estimators, and the numerical results of a computer experiment.

Overall evaluation: *On the positive side the book is an attractive one both in its material and its production. The topics dealt with are well chosen and are described with a minimum of detail. Each chapter is adequately provided with exercises. The publishers have done a good job and the book is nicely produced, particularly the hardback version. For the courses it is intended to cover the book may well be popular succeeding in its stated aims. On the negative side the book could be criticized for a tendency towards an abstract*

mathematical style of presentation with few diagrams or graphs and insufficient illustrations of some of the theoretical ideas. The more straightforward topics of the core syllabus of the book are well explained but some of the additional topics are written in a manner which is difficult to follow. Another criticism which can be aimed at the book is that it skims over basic standard material in favour of pursuing novelty of approach. Thus the treatment of the "bread-and-butter" two-dimensional Gaussian distribution is brief and sketchy whereas the multidimensional Gaussian distribution is treated in some detail including an unusual extended calculation of the inverse of the bivariate covariance matrix. Similarly, ordinary statistical procedures are replaced by new methods using the diagonalization algorithm of Appendix 1. Since the student will need to know the simple unsophisticated procedures later in his or her career, the wisdom of following unusual or nonstandard methods in teaching degree courses is open to question.

The style and philosophy of the book reflects in some measure current trends in teaching. It is quite usually considered that the amount of material which a student might ideally be required to be familiar with nowadays is so extensive that limitations of time (and the students' patience) make it impossible to cover everything in a systematic step-by-step way. So there is a tendency towards a less detailed, informal presentation aimed at touching on recent developments. This runs the risk of not giving the thorough grounding in fundamentals which is essential for good later work. In the reviewer's opinion it is important to try to achieve a synthesis of the clear and systematic presentation of the best of the older books with the livelier presentation of the more recent ones.

About the reviewer

After finishing a first degree in mathematics at Cambridge University, J. F. Barrett joined the then newly-formed control group, graduating with a Ph.D. in 1958 with a thesis on the analysis of randomly disturbed automatic control systems. After appointments at the Universities of Southampton, Addis Ababa and Birmingham, he returned to Cambridge in 1970–1976. After further research appointments in control at the Technical University of Eindhoven, Sheffield Polytechnic and Strathclyde University, he taught mathematics at Paisley College of Technology from where he has recently retired. His main research contributions have been in Volterra series and in Wiener–Kalman filtering and control theory.

References

- Bochner (1932). *Vorlesungen über Fouriersche Integrale*. Teubner, Leipzig.
- Carson, J. R. (1931). The statistical energy-frequency spectrum of random disturbances. *Bell Syst. Tech. J.*, **10**, 374–381.
- Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700–725.
- Gabor D. (1946) Theory of Communication. *J.I.E.E.*, **93** pt 3, 429–441.
- Helstrom, C. W. (1960). *Statistical Theory of Signal Detection*. Pergamon Press, Oxford.
- Landau, H. and H. Pollock (1961). Prolate spheroidal wave functions, Fourier analysis and uncertainty—II. *Bell Syst. Tech. J.*, **40**, 65–84.

Biographical Notes on Contributors to this Issue



Laurent Baratchart was born in 1955. He received his degree in engineering from the Ecole des Mines de Saint-Etienne in 1978, and his doctorat d'état in mathematics from the University of Nice in 1987. His main research interests are nonlinear approximation theory and dynamical systems.

From 1986 to 1988, he worked in the Renault research center on design methods for manufacturing systems. His research interests are in linear multivariable control and performance evaluation of production systems.



Kantibhusan Datta received his B.Sc. (Hons.) in physics, and M.Tech. and Ph.D. in applied physics all from Calcutta University in 1960, 1963 and 1972 respectively. At present he is Professor of Electrical Engineering at IIT, Kharagpur. During 1965-1982, he was attached to the Department of Applied Physics, Calcutta University first as a Lecturer and then as a Reader.

Dr Datta visited the Institute of Mathematics, Bratislava, Czechoslovakia during 1975-1976 and the Department of Mathematics, Eindhoven University of Technology, Holland during 1979-1980 as a Visiting Research Fellow. He served at the Institute of Control Engineering, National Chiao Tung University, Taiwan as a Visiting Expert during 1980-1982. He is a recipient of S. K. Mitra Memorial award (1974) from the Institute of Electronics and Telecommunication Engineers, India. He is a Fellow of the IETE and Institute of Engineers, India and a senior member of IEEE. His fields of interest are multivariable control systems, system identification and applied mathematics.



Michel Cardelli was born in 1962 in Chinon, France. He graduated in applied mathematics from the University of Nice in 1985. He is currently a Ph.D. student at this University. His main research interests are approximation theory and numerical analysis.



Ben M. Chen was born in Fujian, China, on 25 November 1963. He received the B.S. degree in mathematics and computer science from Amoy University, Xiamen, China, in July 1983 and the M.S. degree in electrical engineering from Gonzaga University, Spokane, Washington, in May 1988.

From July 1983 to March 1986 he worked as a software engineer in the South China Computer Corporation, Guangzhou, China. From September 1986 to May 1988 he held a Presidential scholarship at Gonzaga University and a Cardinal Yu-Pin scholarship from the Sino-American Amity Fund, Inc., New York. He is currently working toward his Ph.D. degree in the Department of Electrical and Computer Engineering at Washington State University, Pullman. His current research interests are in robust control theory.



Jean-Michel Dion was born in La Tronche, France, in 1950. He graduated in mathematics in 1972. He received the Thèse de 3ème cycle and Thèse d'Etat degrees, both from the Institut National Polytechnique de Grenoble, in 1977 and 1983 respectively. Since 1979, he has been a researcher at the Centre National de la Recherche Scientifique where he is presently Directeur

de Recherche and vice head of the Laboratoire d'Automatique de Grenoble. He is author or co-author of over 80 journal or conference papers. His current research interests are in linear systems and adaptive control.



Christian Commault received the degree of electrical engineer, the Docteur-Ingénieur degree and the Docteur d'Etat degree from the Institut National Polytechnique de Grenoble in 1973, 1978 and 1983 respectively. From 1974 to 1976 he taught in the Dakar Institute of Technology (Sénégal). Since 1979 he has taught automatic control in the Ecole Nationale Supérieure

d'Ingénieurs Electriciens de Grenoble.

In 1978, he spent one year as a visiting researcher in the Mathematics Institute of Groningen (The Netherlands).



Luc Dugard was born in Vitry-le-François, not far from the famous Champagne vineyards, France. He received the engineer degree in electronics in 1975 from the Institut National Polytechnique de Grenoble. He got his Thèse de Docteur-Ingénieur degree in 1980 and his Thèse de Docteur d'Etat es Sciences degree in 1984, both from the Institut National Polytechnique de Grenoble. Since

1977, he has been with the Laboratoire d'Automatique de Grenoble, E.N.S.I.E.G., where he holds a researcher position at the C.N.R.S. (the French National Center for Scientific Research). His main scientific interests are in the field of adaptive control: theoretical and methodological aspects, and applications to robotics and thermal processes. He has also other interests, and some expertise in, among others, spoonerisms.



Wei ming Feng was born in Nanning, China. She received her first degree from the Dept of Control Engineering, Beijing University of Aeronautics and Astronautics in 1984. She started her research in control engineering in 1985 as a PhD student at Strathclyde University, Scotland, U.K., and later worked as a research engineer in the National Engineering Laboratory, U.K. in 1989. Currently

she is a research associate with Control Systems Research, University of Leicester. Her research interests lie in linear system theory, control systems robustness and control design of industrial robots.



Fouad Giri received the degree of engineer in electrical engineering in 1982, the Doctorat d'Etat in automatic control in 1988 (both from the Ecole Mohammadia d'Ingénieurs in Rabat, Morocco) and the Doctorat in automatic control and signal processing in 1988 from the Institut National Polytechnique de Grenoble, France. From 1982 to 1986, he was an assistant professor in the

Ecole Mohammadia d'Ingénieurs. From 1986 to 1988, he was a researcher in the Laboratoire d'Automatique de Grenoble. Since 1988 he has been a lecturer of Automatic Control at the Ecole Mohammadia d'Ingénieurs. His research interests are in adaptive and robust control.



Kenneth J. Hunt is from Glasgow, Scotland. He obtained a BSc in Electrical Engineering in 1984 and a PhD in Control Theory in 1987, both from the University of Strathclyde, Glasgow. From 1987-1989 he was a Scientist with BBN Systems and Technologies (the European division of Bolt, Beranek and Newman Inc.) in Edinburgh; he worked on expert systems for

process monitoring, intelligent decision support systems for manufacturing applications, and on parallel processing.

Dr Hunt is currently supported by a Personal Research Fellowship of the Royal Society of Edinburgh and is a member of the Control Group in the Department of Mechanical Engineering at the University of Glasgow.

His current research areas include self-tuning control, the polynomial equation approach to optimal control and the application of connectionist architectures to non-linear control problems. Dr Hunt is the author of the monograph *Stochastic Optimal Control Theory with Application in*

Self-tuning Control (Springer, Berlin, 1989). He is a member of the Institution of Electrical Engineers and serves on the Professional Group Committee on Control and Systems Theory.



Ioannis Kanelakopoulos was born in Athens, Greece, in 1964. He received the Diploma in electrical engineering from the National Technical University of Athens in 1987, and the M.S. degree in electrical engineering from the University of Illinois, Urbana, in 1989.

In 1990 he was the recipient of a University of Illinois Grainger fellowship. Since 1987 he has been a Research Assistant at the Coordinated Science Laboratory in Urbana. He is currently working towards the Ph.D. degree in electrical engineering at the University of Illinois, Urbana. His research interests are in the area of adaptive control of nonlinear systems.



Miroslav Kárný was born in Prague, Czechoslovakia in 1948. He received his engineering degree from the Czech Technical University (Faculty of Nuclear and Physical Engineering) Prague in 1973. In 1978, he received his Ph.D. degree from the Institute of Information Theory and Automation, Czechoslovak Academy of Sciences. Since then, he has been with the Institute, currently as senior researcher. His research interests are directed to the theory and industrial applications of self-tuning control.



Pramod P. Khargonekar received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay in 1977, and the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering, from the University of Florida in 1980 and 1981, respectively. From 1981 to 1984, he was with the Department of Electrical Engineering, University of Florida,

and from 1984 to 1989 he was with the Department of Electrical Engineering, University of Minnesota. In September 1989, he joined the University of Michigan where he is a Professor of Electrical Engineering and Computer Science. His current research interests include robust and H_2/H_∞ optimal control, robust adaptive control, distributed systems, time-varying systems and applications to aerospace control problems.

Dr Khargonekar is the recipient of the Donald Eckman award, the NSF Presidential Young Investigator award, and the George Taylor award. He was an associate editor of the *IEEE Transactions on Automatic Control* during 1987-1989. He is currently an associate editor of *Mathematics of Control, Signals and Systems* and *Systems and Control Letters*.



Petar V. Kokotovic received graduate degrees in 1962 in Belgrade, Yugoslavia, and in 1965 in Moscow, U.S.S.R.

From 1959 to 1966 he was with the Pupin Research Institute, Belgrade, Yugoslavia. Since 1966 he has been with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, where he has

supervised 25 Ph.D. students. With them, he has co-authored several books and numerous papers on sensitivity analysis, singular perturbations, large scale systems and adaptive control. He has held visiting appointments with research institutions in the United States, France, Italy, Switzerland and Australia. His industrial consulting activities include those with the Ford Motor Company and the General Electric Company.

Dr Kokotovic is a Fellow of the IEEE, and has served on the Board of Governors and IDC of the Control Systems Society, on committees of the International Federation of Automatic Control (IFAC), and as an Associate Editor of several technical journals.



John J. Kornylo received the B.S. degree in engineering science from Florida State University, Tallahassee, in 1972, the M.S. degree in engineering science from the University of South Florida, Tampa, in 1976, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1979. He is currently a Research Specialist at the Exxon

Production Research Company, Houston, TX. His interests lie principally in numerical methods for estimation and optimization. Dr Kornylo is a member of the IEEE, the Society of Exploration Geophysicists, Eta Kappa Nu, and Tau Beta Pi, among others. His hobbies include duplicate bridge, tennis, computer games and libertarian politics.



Alexander H. Levis was born in Yannina, Greece, in 1940. He was educated at the Massachusetts Institute of Technology where he received the BS (in 1965), MS (in 1965), ME (in 1967), and Sc.D. (in 1968) degrees in mechanical engineering. He also attended Ripon College in Wisconsin where he received the AB degree (1963) in Mathematics and Physics. From 1968 to

1973 he was on the Faculty of the Department of Electrical Engineering of the Polytechnic Institute of New York, Brooklyn (now Polytechnic University). From 1973 to 1979, he was with Systems Control, Inc. of Palo Alto, CA, where his last position was manager of the Systems Research Department. Since 1979, Dr Levis has been a Senior Research Scientist at the MIT Laboratory for Information and Decision Systems. Dr Levis' current research interests include the modeling, analysis and evaluation of architectures for distributed intelligence systems as well as the modeling of human decision-making in information processing organizations. He is a Fellow of IEEE and past president of the IEEE Control Systems Society. He has served as an editor of *Automatica*, as chairman of the IFAC Technical

Committee on Systems Engineering and is currently serving as vice-chairman of the IFAC Technical Board.



F. L. Lewis studied in Chile and Scotland, subsequently obtaining a MEE degree at Rice University in Houston. He spent six years in the U.S. Navy, serving as Navigator, Executive Officer, and Acting Commanding Officer aboard USS Salinan ATF-161. In 1981 he obtained the Ph.D. degree at Georgia Institute of Technology in Atlanta, where he is now a professor. He has

studied the geometric properties of the Riccati equation, and his current interests include singular systems and robotics. He is the author of *Optimal Control and Optimal Estimation*. Dr Lewis is a Senior Member of the IEEE, Associate Editor of *Circuits, Systems and Signal Processing*, and the recipient of a Fulbright Award, the ASFE Terman Award, and two Sigma Xi Research Awards.



Riccardo Marino was born in Ferrara, Italy, in 1956. He received his degree in nuclear engineering in 1979 and a Masters degree in systems engineering in 1981 from the University of Rome "La Sapienza". In 1982 he received the Doctor of Science degree in systems science and mathematics from Washington University, St. Louis, Missouri, U.S.A. Since 1984 he has been

with the Department of Electronic Engineering at the University of Rome "Tor Vergata", where he is currently Associate Professor in Systems Theory. In 1986 he visited the University of Twente, Enschede, The Netherlands and in 1985 and 1989 the University of Illinois in Urbana-Champaign. His research interests are in nonlinear control theory.



James S. McDonald was born in El Dorado, Arkansas, in 1957. He received the degrees of S.B.E.E. and S.M.E.E. from the Massachusetts Institute of Technology, Cambridge, in 1980. His Master's thesis research was carried out while he was a student employee at the General Electric Company Corporate Research and Development Center, Schenectady, New York. In 1981 and

1982 he was an Instructor in the Department of Electrical Engineering at the University of New Orleans, LA, where he was recognized as the Outstanding Teacher in Electrical Engineering for the 1981-82 academic year. From 1982 until 1987 he was with Lawrence Livermore National Laboratory, Livermore, CA, first in the Beam Research Program and, from 1985, in the Dynamics and Controls Group of the Engineering Research Division. From 1985 to 1987 he was also a graduate student in the Department of Electrical Engineering at Stanford University, CA. Since 1988 he has been a research assistant in the Department of Electrical and Computer Engineering at Rice University, Houston, TX, where he is working toward the Ph.D. degree. His main research interests are in optimal and robust control of linear multivariable systems.



Richard H. Middleton was born in Newcastle, Australia, in 1961. He obtained a B.Sc. (physics), B.E. (electrical engineering) and Ph.D. from the University of Newcastle, Australia. Since 1986 he has been a lecturer in the Department of Electrical Engineering and Computer Science at the University of Newcastle, and has recently spent a sabbatical leave at the Co-ordinated Science

Laboratory, University of Illinois at Urbana-Champaign. He is co-author of the book *Digital Control and Estimation* (Prentice Hall, Englewood Cliffs, NJ, 1990). Dr Middleton is a member of the IEEE, and his interests include adaptive control, digital control, satellite tracking systems and industrial electronics.



Mario Milanese was born in Alessandria, Italy, on 11 September 1942. He received the laurea degree in electronic engineering from the Politecnico di Torino in 1967. From 1967 to 1980 he was Assistant Professor at the Politecnico di Torino and from 1972 to 1981 Associate Professor of System Theory at the University of Turin. Since 1980 he has been full Professor of System Theory at

the Politecnico di Torino, where from 1982 to 1987 he was Chairman of the Dipartimento di Automatica e Informatica. Dr Milanese is at present member of GNASI (Gruppo Nazionale di Automatica Sistemistica e Informatica) and associate editor of *Information and Decision Technologies* (North Holland). His research interests include modeling, robust identification and control, forecasting and decision support systems.



Pradeep Misra received the B. Tech. degree in electrical engineering from Indian Institute of Technology, Kharagpur, India, and the Ph.D. degree from Concordia University, Montreal, P.Q., Canada.

Since 1987, he has been an assistant professor of Electrical Engineering at Wright State University, Dayton, OH, U.S.A. and an Adjunct Assistant Professor of

Electrical and Computer Engineering at Concordia University. His research interests are in computational methods for analysis and design of high-order systems, modelling and control of robot manipulators and multi-dimensional digital signal processing.



Bosukonda Murali Mohan was born in Tapeswaram, India, in 1960. He received his B.E. in electrical engineering from Osmania University in 1982, and his M.E. and Ph.D. in control systems from the Department of Electrical Engineering at Andhra University, Visakhapatnam, and the Indian Institute of Technology, Kharagpur in 1985 and 1990, respectively. During 1989-90, he

was a lecturer in the Department of Electrical and Electronics Engineering at Regional Engineering College, Tiruchirapalli, India. Presently he is a lecturer in electrical engineering at IIT, Kharagpur. His research interests are in system theory, system identification, parameter estimation and applications of orthogonal functions in signals, systems, and control.



Mohammed M'Saad was born in Angads-Oujda, Morocco, in 1953. He graduated from the Ecole Mohammadia d'Ingénieurs, Rabat, Morocco, in 1978 as an electrical engineer. He obtained the degree of Docteur de 3ème cycle from the Faculté des Sciences, Rabat, Morocco, and Docteur d'Etat from the Institut National Polytechnique de Grenoble, France, in 1982 and 1987,

respectively. He was Maître-Assistant and Maître de Conférence at the Ecole Mohammadia d'Ingénieurs and researcher at the Laboratoire d'Electronique et d'Etude des Systèmes Automatiques. He is currently researcher at the Centre National de la Recherche Scientifique (C.N.R.S), France. His research interests are in adaptive control.



Martine Olivi was born in Marseille, France, in 1958. She received the engineer degree from the School of Mines of St-Etienne, France, in 1983, and the thesis in mathematics from the University of Marseille, in 1987. Since 1986 she has been with the INRIA, Sophia-Antipolis, where she presently holds a Researcher position. Her main research interests are in

rational approximation theory.



Hitay Özbay was born in Ankara, Turkey, on 17 May 1962. He received the B.S. degree from the Middle East Technical University, Ankara, Turkey, in 1985, and the M.Eng. degree from McGill University, Montreal, Canada, in 1987, both in electrical engineering. After receiving the Ph.D. degree from the Control Sciences and Dynamical Systems Center of the University of Minnesota, Dr

Özbay joined the Electrical Engineering Department of the University of Rhode Island, as an Assistant Professor, in 1989. His research interests include H^∞ optimal control, robust control, distributed parameter systems, system identification and adaptive control. Dr Özbay is a member of IEEE.



Rajnikant V. Patel received the B.Eng. degree in electronics from the University of Liverpool, U.K., in 1969 and the Ph.D. degree in electrical engineering from the University of Cambridge, U.K., in 1973.

He has held postdoctoral and visiting faculty positions at the University of Cambridge, U.K., Lund Institute of Technology, Sweden, NASA Ames Research

Center, U.S.A., University of Waterloo, Canada, and Delft University of Technology, Holland. He was a lecturer in Control Engineering at the Control Systems Centre, UMIST, England, from 1980 to 1981. At present, he is a Professor of Electrical and Computer Engineering at Concordia University, Montreal, Canada. His current research is concerned with various theoretical and computational issues in robotics and control. He has also conducted research on computer-aided control system design, structure and design of multivariable systems, robustness in estimation and control, adaptive stochastic control and computational methods for high-order systems. He has published numerous technical papers in these areas and is the author (with N. Munro) of the book, *Multivariable System Theory and Design* (Pergamon, Oxford, 1982). He is an Associate Editor of the *IEEE Transactions on Automatic Control* and of *Automatica*.



J. B. Pearson received the Ph.D. degree from Purdue University, West Lafayette, IN, in 1962. He served as Assistant Professor of Electrical Engineering at Purdue from 1962 to 1965, and then joined the Department of Electrical Engineering at Rice University, Houston, TX, where is now J.S. Abercrombie Professor in Engineering. He served as Associate Editor for the *IEEE*

Transactions on Automatic Control from 1975 to 1977 and again from 1986 to 1988. He was president of the IEEE Control Systems Society in 1984. In 1985, he received with B.-C. Chang an Outstanding Paper award and in 1989, with M. A. Dahleh, a George S. Axelby Outstanding Paper award, both from the Control Systems Society.



Didier M. Perdu is a research Engineer at Architecture Systèmes Avancés, a subsidiary of Thomson-CSF. Prior to this, he was a visiting scientist at the Massachusetts Institute of Technology Laboratory for Information and Decision Systems. Born in Paris, France in 1962, he graduated from the Ecole Supérieure d'Electricité (1986) and received the M.S. degree in Technology

and Policy from the Electrical Engineering and Computer Science Department of MIT (1988). His research interests are in decision aids and modeling and evaluation of C3I systems.



Johannes Prock was born in Berlin, (West) Germany, on 22 June 1955. He received the engineering degree and the Ph.D. degree in electrical engineering, both from the Technical University of Munich, Germany, in 1981 and 1986 respectively. Since then he has been with the Gesellschaft für Reaktorsicherheit, Garching, Germany, where he is currently leading a project concerning signal

validation in nuclear power plants. His research interests lie mainly in the area of low-order accurate process models, qualitative models, failure detection methods and microcomputer and transputer applications.



Mario A. Rotea was born in Rosario, Argentina, on 6 August 1958. He received the degree of electronic engineer from the National University of Rosario, Argentina in 1983. He obtained the M.S. degree in electrical engineering and the Ph.D. degree in control science and dynamical systems from the University of Minnesota in 1988 and 1990, respectively. In 1989 he was

awarded a Doctoral Dissertation Fellowship by the Graduate School of the University of Minnesota.

From 1983 to 1984, he was an Assistant Engineer at the Military Ammunition Factory "Fray Luis Beltrán", Argentina. From 1984 to 1986 he was a Research Associate at the Institute of Technological Development for the Chemical Industry, Santa Fe, Argentina. Currently, Dr Rotea is an Assistant Professor in the School of Aeronautics and Astronautics at Purdue University. His research interests include robust multivariable control, optimal control and applications of control theory to aerospace problems.



Ali Saberi received his Ph.D. degree in electrical engineering from Michigan State University, East Lansing, in 1983.

He is currently an Associate Professor in the Electrical and Computer Engineering, Washington State University, Pullman, Washington.



Peddapulliah Sannuti was born in Rajupalem, Proddatur (Taluk), India. He received the B.E. degree (with honours) in electrical engineering from the College of Engineering, Anantapur, India, in 1963, the M.Tech. degree in control systems engineering from the Indian Institute of Technology, Kharagpur, India, in 1965, and the Ph.D. degree in electrical engineering from the

University of Illinois, Urbana-Champaign, in 1968.

From 1965 to 1968, he was firstly a Teaching Assistant and then a Research Assistant in the Department of Electrical Engineering and the Coordinated Science Laboratory, University of Illinois. Since 1968, he has been with the

Department of Electrical Engineering, Rutgers University, Piscataway, NJ, where he is currently a Professor. His research interests include singular perturbation theory, computational methods and communication theory.



Michael Šebek was born in Prague, Czechoslovakia in 1954. He received the Ing. degree in electrical engineering from the Czech Technical University, Prague in 1978 and the CSc (Ph.D.) degree in control theory from the Czechoslovak Academy of Sciences in 1981. Since 1981 he has been with the Institute of Information Theory and Automation, Czechoslovak Academy

of Sciences, Prague. He has held visiting positions at the University of Padova, Italy, at the University of Strathclyde, Glasgow, Scotland and at the University of Toronto, Canada. He is currently on leave at the Department of Applied Mathematics, University of Twente, Netherlands. He was awarded the Young Scientist Award from the Czechoslovak Academy of Sciences in 1987 and the Czech National Prize from the Czech Parliament in 1989. His research interests are in linear systems theory including n -D and delay-differential systems, algebraic control theory and robust control including H_∞ -optimization.



Malur K. Sundareshan received the B. E. degree in electrical engineering from Bangalore University, in 1966, and the M. E. and Ph.D. degrees in electrical engineering from the Indian Institute of Science, Bangalore, India in 1969 and 1973, respectively.

Between 1973 and 1976, he held various visiting faculty positions at the Indian Institute of

Science, Bangalore; at the University of Santa Clara, California; and at Concordia University, Montreal, PQ, Canada. From 1976 to 1981, he was on the faculty of the Department of Electrical Engineering, University of Minnesota, Minneapolis. Since 1981, he has been on the faculty of the Department of Electrical and Computer Engineering, University of Arizona, Tucson, where he is a professor. His current research interests are in large scale systems, communication networks, adaptive control and estimation, and statistical signal processing.



Vassilis L. Syrmos received the Diploma in electrical engineering from the Democritus University of Thrace, Xanthi, Greece, in 1987. Since then he has been a Research Assistant at the Georgia Institute of Technology, Atlanta, where he is currently working towards the Ph.D. degree in electrical engineering. His research interests lie in the areas of geometric theory and matrix pen-

cil theory, and their applications to linear system theory. He is the recipient of the AHEPA Scholarship and the A. S. Onassis Fellowship and is a member of Pi Mu Epsilon and associate member of Sigma Xi.



Allen Tannenbaum was born in New York City in 1953. He attended Columbia and Harvard Universities where he received his Ph.D. in mathematics in 1976. He won the attendance medal from Benjamin Cardozo Junior High School in 1966, and came in third in the quarter mile in the New York City high school championships in 1969. (His time was 50.5 seconds.) He has also served as a

combat medic in the Israeli army. Dr Tannenbaum presently is teaching at the Technion in Israel and at the University of Minnesota. His research interests are in robust control, operator theory, and computer vision.



Jorge A. Torres Muñoz was born in Mexico City in 1960. He received the B.S. degree in electronic engineering from the National Polytechnic Institute of Mexico in 1983, the M.S. degree in electrical engineering from the Centro de Investigación y de Estudios Avanzados (CIEA) of Mexico in 1985 and the Ph.D. degree in automatic control from the Polytechnic Institute of Gren-

oble, France in 1990. From 1985 to 1986 he was a research assistant at CIEA of Mexico. He is presently as associate professor at CIEA-Mexico. His main interests are in the geometric and algebraic approaches to linear systems.



Antonio Vicino received the Laurea in electrical engineering from the Politecnico di Torino, Italy, in 1978. From 1979 to 1982 he held several Fellowships at the Dipartimento di Automatica e Informatica of the Politecnico di Torino. He was Researcher of Automatic Control from 1983 to 1987 at the same Department. In 1987 he joined the Dipartimento di Sistemi e Informatica,

Università di Firenze, Italy, as Associate Professor of Automatic Control. Presently, he is Professor of Automatic Control and System Theory. His research activities are mainly in the fields of robust stability and control, applied system modelling and time series prediction, robust identification. Prof. Vicino is a member of GRIS (Gruppo Ricercatori Informatica e Sistemistica).

A Practical Study of Adaptive Control of an Alumina Calciner*

P. M. MILLS,† P. L. LEE‡ and P. McINTOSH†

An adaptive pole-placement controller has been applied to an industrial alumina calcination process, and it has been found that this control method has reduced temperature variations by three fold and reduced overall energy consumption.

Key Words—Adaptive control; alumina processing; calcination; pole-placement.

Abstract—This paper outlines the successful application of Advanced Process Control methods in an industrial environment. The control strategy incorporated conventional regulatory control techniques, dead-time compensation and an explicit pole-placement self-tuning control algorithm. Application of this strategy to control an alumina calcining kiln is successfully demonstrated. Success is demonstrated not only by improved regulatory behaviour, but in terms of operator and engineering acceptance and on-going use of the control strategy.

1. INTRODUCTION

THE NEED to improve efficiency of process industries is well recognized. Recent studies (Marlin *et al.*, 1987) have shown that application of Advanced Process Control is one available tool to “make more with less” with good economic return. Typical returns have been in the order of 5–10% of annual production value. Availability of modern distributed control systems, often in conjunction with process computer systems, has made technical application of advanced control methods feasible. However, process industries in general have been slow to utilize this increased capability.

Adaptive control has probably been the topic most written about, academically, in control literature in the past 15 years. The field is rich with different approaches for parameter identification, controller design, maintaining signal excitation and numerical techniques. However,

despite this richness, few industrial applications have been reported in the process industries that have passed the so-called “12 month test” i.e. the controller is still operating 12 months after being commissioned and is being maintained by normal operating and instrument personnel—not the development team. A review of published applications was given by Seborg *et al.* (1986). The ability of an adaptive controller to adapt to changing process behaviour, particularly changes caused by the underlying nonlinearity of many processes, makes the adaptive approach appealing. However, so few successful industrial applications are reported in the open literature that the processing industries are collectively “nervous” about applying such “modern” techniques to their plants. Only the economic incentives mentioned previously and reporting of successful applications will change this situation.

This paper describes one such successful industrial application of adaptive control. It highlights some practical issues in application of such methods in an industrial environment and reports the economic improvement obtained in using such approaches. This paper does not provide a new theoretical insight but rather describes a successful application of academically established techniques.

2. THE APPLICATION

Queensland Alumina Limited (QAL) operates an alumina refinery at Gladstone, Queensland based on the Bayer process. Maximum plant capacity is 3,000,000 metric tonnes of alumina per year. The process extracts alumina from bauxite by dissolution in caustic soda. The final phase of production is the drying and calcining,

* Received 22 May 1989; revised 2 May 1990; received in final form 4 September 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor L. Keviczky under the direction of Editor H. Austin Spang, III.

† Queensland Alumina Ltd, Gladstone, QLD 4680, Australia.

‡ Process Control Group, Department of Chemical Engineering, University of Queensland, St. Lucia, QLD 4067. Author to whom all correspondence should be addressed.

- (c) The interaction between oxygen and cold-end temperature loops which was relieved by special tuning techniques, forces any re-tuning to be done on a kiln basis rather than individual loops. This made initial tuning and tuning maintenance time-consuming.

In 1987 work began on development of an advanced control strategy to overcome these deficiencies. The kiln chosen for this study is a GATX-Fuller rotary kiln of 3.66 m diameter, 99 m long, generally rotating at a fixed speed of 1 r.p.m. The kiln capacity is approximately 1000 tonne per day. Heat recovery exists only from the product at the hot end, not on the flue gas at the cold end.

The general structure of the advanced control strategy is shown in Fig. 2. The important features of this strategy are:

- The cascade CO-O₂ control strategy to maintain combustion efficiency. The inner O₂ controller includes a Smith Predictor (1957) for compensating the inherent dead-time in this loop.
- A self-tuning pole-placement hot-end temperature controller adjusting hydrate feed rate. This controller is discussed in more detail below.
- Operator-set oil mass-flow adjustment which effectively allows the operator to set throughput for the kiln. It was found necessary to adopt this strategy rather than

allowing the operator to set the kiln feedrate directly and controlling the HET by the oil flowrate. This was done because complex interactions within the kiln can cause the open-loop gain of a HET-oil loop to change sign as a function of operating conditions. This obviously creates serious problems for any control strategy employing this configuration. However, the open-loop gain between the oil flowrate and the peak temperature attained in the kiln (and hence LOI) is always positive, always allowing the operator to increase the hydrate setpoint following an increase in the oil flowrate.

- (d) Feed-forward compensation from both adjustments made in the kiln feedrate and oil flowrate to the draft controller.

3.1. Self-tuning controller

An explicit pole-placement self-tuning controller as developed by Wellstead *et al.* (1979) was used for the HET controller. This controller uses as its basis a process model of the form:

$$(1 + A(Z^{-1}))y(t) = B(Z^{-1})u(t) + e(t) \quad (1)$$

where $A(\)$ and $B(\)$ are polynomials in the backward shift operator Z^{-1} of order a and b respectively, $y(t)$ is the error in the HET (setpoint-measurement), $u(t)$ is the process input (kiln feed-rate), $e(t)$ is the model error.

Dead-time in the above process model is included by increasing the order of the B polynomial and setting the appropriate leading coefficients to zero. The regulator takes the form:

$$(1 + F(Z^{-1}))u(t) = G(Z^{-1})y(t) \quad (2)$$

where polynomials F and G are of order f and g respectively.

The pole-placement algorithm requires selection of F and G such that the closed-loop pole locations are located to desired values as specified by the polynomial

$$1 + T(Z^{-1}) \quad (3)$$

where T is of order t . By combining equations (1), (2) and (3), the following identity is obtained:

$$\{1 + A(Z^{-1})\}\{1 + F(Z^{-1})\} - B(Z^{-1})G(Z^{-1}) = 1 + T(Z^{-1}).$$

Important constraints are

$$g = a - 1$$

$$f = b - 1$$

$$t < a + b - 1.$$

A solution of the above identity given A , B and

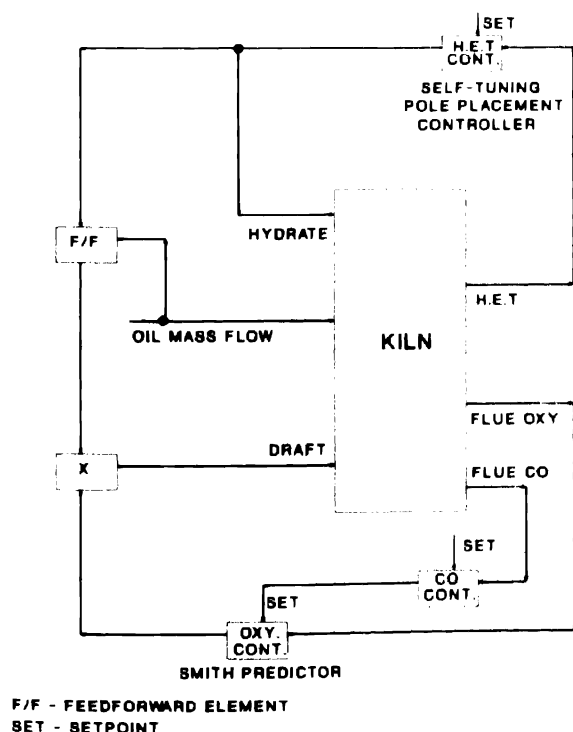


FIG. 2. Revised control scheme.

T polynomials is required at each sampling interval. A matrix approach to this problem is outlined by Wellstead and Prager (1985). Gauss Jordan elimination was used in this application and proved sufficiently reliable.

Controller design was achieved by specifying the closed-loop bandwidth (B) and relative damping (D). Given these two specifications, Wittenmark and Åström (1984) have shown that the order of the T polynomial is 2, and that:

$$t_1 = -2e^{-D\pi} \cos(B(1 - D^2)^{1/2})$$

and

$$t_2 = e^{-2D\pi}.$$

A damping factor of approximately 0.7 and bandwidths between 0.25 and 5 radians/sample proved satisfactory for this application, with a bandwidth of 1.5 radians/sample finally chosen.

Recursive least-squares parameter identification with a forgetting factor was employed to provide on-line estimates of the A and B polynomial parameters. The UD factorization method (Bierman, 1976) was used in the co-variance update to improve numerical accuracy.

3.2. Implementation issues for the self-tuning controller

Implementing a self-tuning controller in an industrial environment raises a number of practical issues. Among these are:

(a) *Hardware implementation strategy.* For reasons dominated by reliability it was decided to implement the control law using a sampled data control algorithm available in the Bailey Network 90 system, and the adaptive strategy on the Data General MV8000. This also allows the adaptive strategy to adapt several loops simultaneously using the same software.

(b) *Sample time.* Choice of an appropriate sampling time is governed by the dynamics of the process and the desire to keep the number of coefficients in the identified model suitably small—approximately 5 for each polynomial. Six minutes proved suitable for this application.

(c) *Sample filter.* The procedure of sampling a continuous signal can introduce problems of aliasing as discussed by Åström and Wittenmark (1984). Careful consideration of the process and measurement system must be taken into account in designing an anti-alias filter. For example, in this application it was determined that the HET thermocouple probe mounting contributed an equivalent first-order lag of two minutes. The required time constant for an anti-alias filter for a sample time of six minutes is approximately

four minutes. Hence the required additional filter is approximately two minutes.

(d) *Scaling.* Engineering unit ranges of the input and output variable are considerably different. To obtain more reliable parameter estimates, scaling of the differenced HET samples was used to make the ranges nearly the same.

(e) *Model order.* The model order of the A and B polynomials has an important impact on computational load of the identification algorithm, rising as a square relationship with the total number of parameters. Model order is also dependent on the chosen sample time. High order models of course offer the advantage of greater modelling accuracy, and thus a compromise is required. Model orders of 4 and 5 for the polynomials A and B respectively were chosen for this study. These model orders were obtained experimentally from simulation of the regulator with inverse response models (as is the feed/HET transfer function), and during actual kiln trials. The high order of the B polynomial is necessary for the inverse response while the order of the A polynomial combined the actual process order with the external integrator compensation plus the implicit noise model order (Wellstead *et al.*, 1979).

(f) *Model validation.* To provide some protection against an upset-corrupted model and to allow an automated start-up of the regulator, an algorithm which checks for model validity was added. This is an OK/NOT OK check. If NOT OK, the regulator design uses a back-up set of controller parameters. Criteria used for this check were:

$$-1.5 < a_1 + \dots + a_4 < -0.5$$

and

$$0.1 < b_1 + \dots + b_5 < 2.0.$$

The first criteria ensures that the identified model is relatively stable ($\sum a_i \approx 1$) while the second criteria ensures that the steady-state model gain is "reasonable" ($\sum b_i \approx +1$). The back-up controller constants were

$$f_1 = f_2 = \dots = f_4 = 0$$

$$g_0 = \frac{\text{scaling factor}}{5}$$

$$g_1 = g_2 = g_3 = 0.$$

(g) *Forgetting factor.* The exponential forgetting factor used in the recursive least squares algorithm was set to 0.99 for this application. This represented a reasonable compromise between steady-state accuracy and speed of adaptation.

(h) *Signal excitation.* During periods of good

control, little information is available to the identification algorithm. To avoid parameter "explosion", a dither signal was constantly added to the controller output signal. This dither signal was a square wave with constant amplitude but randomly selected period.

(i) *Setpoint rate-of-change*. Large changes in HET setpoint would be observed by the identification algorithm as a sudden unmeasured disturbance with the potential for introducing model errors. As the primary purpose of this controller is to act as a regulator, servo response was limited by placing rate-of-change limits on the setpoint. While a multi-input, single-output identification procedure could have been used to overcome this problem, the performance requirement and the fact that setpoint changes occur infrequently justified this simple approach.

(j) *Pretune*. To enable rapid initialization of the controller and a fast recovery after model

corruption, a rapid pre-tune mode was developed. When the controller operated in this mode, the dither signal amplitude was increased to three times its normal value and the exponential forgetting factor was set to 0.95. The control algorithm used back-up parameters during this phase until the model validation criteria was again satisfied.

The overall controller algorithm structure is shown in Fig. 3.

4 CONTROL PERFORMANCE

4.1. Combustion control

Improved performance of the CO-O₂ control strategy incorporating the dead-time compensator is shown in Fig. 4. This clearly shows reduction in oxygen content in the stack gas achieved because of the higher setpoint on the CO controller. This achievement was due to reduced variation in the oxygen measurement

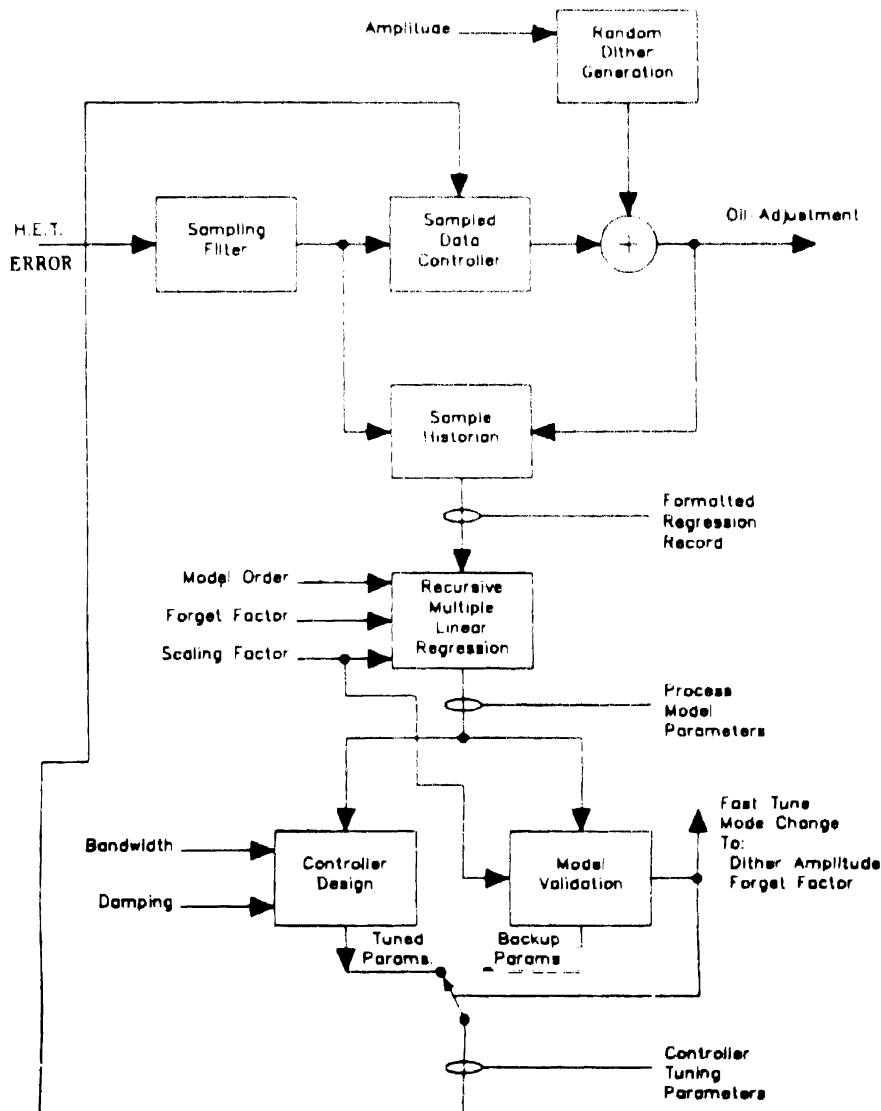


FIG. 3. Pole-placement self-tuning regulator algorithm.

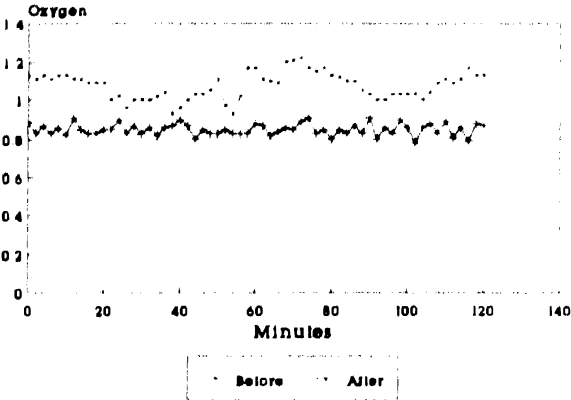


FIG. 4. Normalized oxygen level.

also evident in Fig. 4. The remaining variation with a dominating apparent period of 1 minute is probably caused by the kiln rotation of the fresh and variable moisture hydrate feed. Improvement may be achieved by a slight detuning of this control loop.

4.2. HET control

Improved performance of the HET self-tuning control strategy is shown in Fig. 5 which shows that the histogram of HET has been narrowed and the mean shifted to a more central position. This is confirmed statistically as is shown in Table 1 which also shows that at a 99% confidence level the mean and standard deviation have changed.

The histograms shown in Fig. 5 were established by collecting HET values at five minute intervals for 3 months prior to the installation of the control strategy (26,000 values) and one month after the strategy was implemented (8000 samples). Data was collected continuously during these periods with no distinction made between periods of automatic and manual control. It was estimated that the control strategy was in automatic mode for 95% of the time, with periods in manual control generally attributable to equipment maintenance.

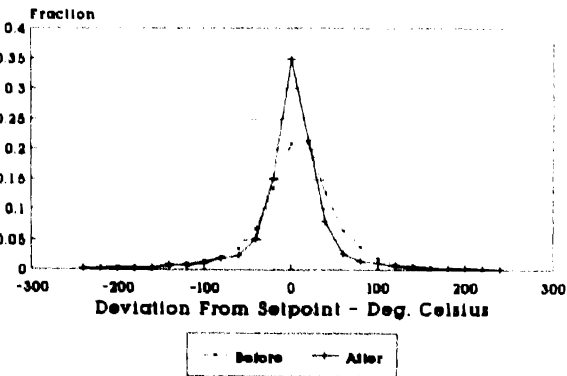


FIG. 5. Hot end temperature histogram—trial kiln.

TABLE 1. IMPROVED HET CONTROL

Characteristic	Before	After	Statistic	99% Confidence index
Mean (°C)	5.9	1.7	$t = 6.13$	2.33
Standard deviation (°C)	55.4	47.4	$F = 1.17$	1.025
% Within +30°C	55.1	71.5		

To confirm that improved control performance could be attributed to the advanced control strategy, a comparison of HET control before and after the new strategy was implemented on the trial kiln was conducted on other kilns of the same type (Fig. 6). This indicates that no external effects can be responsible for the observed improvement.

Other kiln improvements include:

- (a) Improved overall kiln stability leading to the expectation of reduced energy consumption and more consistent product quality.
- (b) The HET regulator is capable of making large prompt hydrate feed changes when feed density alters which prevents subsequent damage to the kiln brickwork.
- (c) The kiln has demonstrated stable operation on automatic control over a much wider range of throughputs down to 2.9 tonne per hour oil flowrate compared to 3.5 tonne per hour previously considered the minimum.

Average oil consumption figures per tonne of hydrate processed could not be obtained as the hydrate measurement shows considerable drift over long periods, making comparisons before and after implementation impossible. This comparison could be performed after *all* kilns were converted to running the new control strategy based upon the total alumina conveyor and inventory figures.

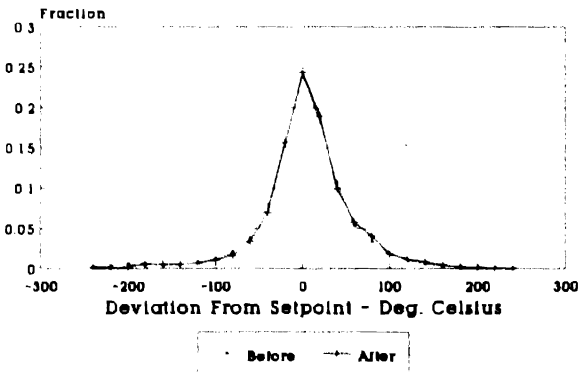


FIG. 6. Hot end temperature histogram—all other kilns.

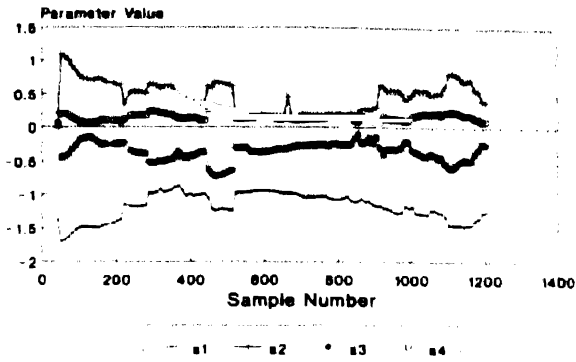


FIG. 7. Normalized model parameter trends (*A* polynomial).

A plot of the normalised identified model parameters are shown in Figs 7 and 8. These show the changes that occur in the *A* and *B* polynomial coefficients over time. It is clear from these figures that the process does indeed change, requiring adaption of the model parameters, clearly justifying the need for an adaptive controller. This is further illustrated in Fig. 9, showing the impulse response derived from the model parameters at sample instants 604 and 884 from Figs 7 and 8. Since the regulator is cascaded with an integrator, the impulse response of the identified model is equivalent to the step response of the actual process dynamics. The model of sample 604 shows insignificant inverse response with between 12 and 18 minutes of dead-time, while the model of sample 884 indicates a 30% peak at the inverse response, and a lower overall gain. Again, this highlights the need for an adaptive controller.

4.3. Human factors

(a) *Control room operators.* Ease of use of the control strategy was important in its overall success. The two new controller types (Smith Predictor and Self-Tuning Regulator) appear to the operator like any other control loop on the Bailey Operator Interface Unit. Acceptance of the strategy increased with familiarity and was

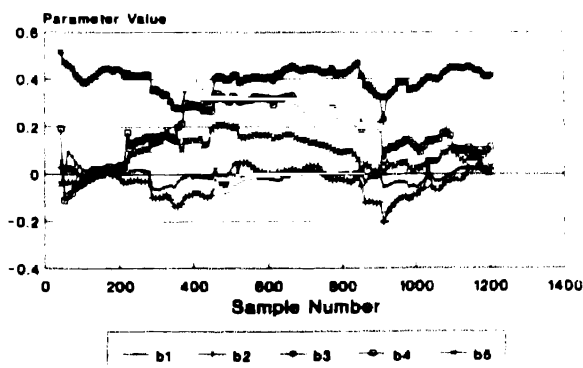


FIG. 8. Normalized model parameter trends (*B* polynomial).

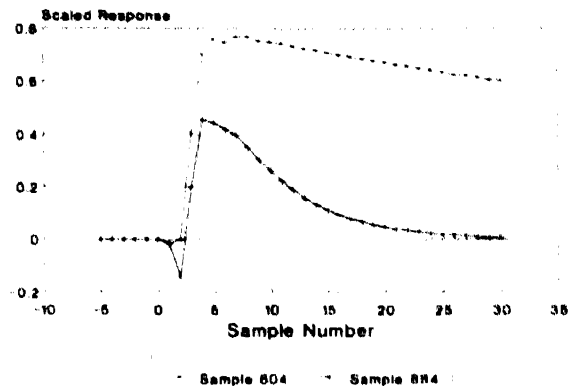


FIG. 9. Impulse model response.

indeed aided by this "sameness" appearance on the visual display unit.

(b) *Instrument personnel.* Although instrument personnel do not have to maintain the strategy, education regarding the strategy's principles was important to prevent it becoming the "scapegoat".

(c) *Engineering personnel.* One engineer was required for design and implementation of the strategy for an estimated period of 12 months. This period also included a learning phase on the principles of self-tuning control. Careful documentation was also important to allow future applications of this technology to proceed.

5. CONCLUSION

This paper has outlined the successful application of advanced control principles in an industrial environment. The combination of conventional controls, a Smith Predictor and an explicit pole-placement self-tuning controller has been shown to be very effective on the control of an alumina calcination kiln. A measure of the success and acceptance of the new control system is that the original trial kiln has now been controlled by the strategy for over 20 months, and two other kilns for over 12 months and all are still operating. There are plans to extend the system to the remaining kilns.

Acknowledgements—The authors greatly appreciate the co-operation of QAL in preparing this paper and the ability to publish such a case study. The help and assistance of many others at QAL is also gratefully acknowledged.

REFERENCES

- Åström, K. J. and B. Wittenmark (1984). *Computer Controlled Systems: Theory and Design*. Prentice-Hall, Englewood Cliffs, NJ.
- Bierman, G. J. (1976). Measurement updating using the U-D factorization. *Automatica*, **12**, 375–382.
- Marlin, T. E., G. W. Barton, M. L. Brisk, and J. D. Perkins (1987). *Advanced Process Control Project Report*, Warren Centre for Advanced Engineering, University of Sydney.
- Mills, P. M. (1988). *Advanced control of the alumina*.

- calcination process. ME Thesis, University of Queensland, Australia.
- Seborg, D. E., T. F. Edgar and S. L. Shah (1986). Adaptive control strategies for process control: A survey. *AIChEJ* **32**, 881–913.
- Smith, O. J. M. (1957). Close control of loops with deadtime. *Chem. Engng Prog.* **53**, 217.
- Wellstead, P. E. and D. Prager (1985). Self-tuning multivariable regulators. In C. J. Harris and S. A. Billings (Eds). *Self-tuning and Adaptive Control: Theory and Applications*. Peter Peregrinus, London.
- Wellstead, P. E., D. Prager and P. Zanker (1979). Pole assignment self-tuning regulator. *Proc. IEE.* **126**, 781–787.
- Wittenmark, B. and K. J. Åström (1984). Practical issues in the implementation of self-tuning control *Automatica* **20**, 595–605.

Adaptive and Robust Cascade Schemes for Thyristor Driven DC-motor Speed Control*

R. M. STEPHAN,[†] V. HAHN,[‡] J. DASTYCH[‡] and H. UNBEHAUEN[‡]

The widely used cascade speed control scheme for thyristor driven DC-motors can be significantly improved by applying adaptive and robust control methods to compensate for unmeasurable variations of their mechanical and electrical characteristics.

Key Words—Adaptive control; robust control; DC-motor; speed control; microcomputer-based control; direct digital control; cascade control.

Abstract—This paper is concerned with the development of improved cascaded speed control systems for thyristor driven DC-motors. The main features of the work can be summarized in the following four points. (1) Development of a digital dual-mode adaptive controller for the inner current control loop and of a model reference adaptive controller for the outer speed control loop, thus making the entire system adaptive. (2) Development of robust controllers both for the inner current loop and also for the outer speed loop, thus making the entire system robust. (3) Implementation of the above control strategies, adaptive and robust, in a 16-bit single board computer with floating-point coprocessor. (4) Comparison of the results of both robust and adaptive improved cascaded schemes with a commercially available controller. The obtained results showed that the model reference adaptive control concept and the robust control strategy can be applied with success for the speed control of a DC-motor.

1. INTRODUCTION

AMONG various methods proposed nowadays for the speed control of thyristor driven DC-motors, cascaded speed control schemes, with an inner current control loop and an outer speed control loop have been widely used and represent the classical control structure.

Certain difficulties arise in actual practice in this system. One is due to the conduction modes of the armature current. As long as the current is continuous, the armature may be modelled

practically as a first order system, but as the current becomes discontinuous, it exhibits nonlinear gain behavior (Kümmel, 1965; Buxbaum, 1969; Pelly, 1971). Furthermore, variations in load torque, moment of inertia of the load and field excitation may also occur. Under these conditions desirable performance of the system may be maintained either by employing adaptive control techniques or by incorporating robust control designs. Usually a fast inner current loop compensates for the armature behavior and a relatively slow outer control loop compensates for the rest.

As the continuous and discontinuous operating modes can be easily detected with inexpensive circuits, it is common practice to apply the so called "parameter scheduling control" principle (Unbehauen, 1985b) for the inner adaptive current control loop. The scheduled adaptation, that will be termed here also as dual-mode adaptation, responds as soon as a plant variation is detected changing the controller parameters correspondingly. This method was introduced by Buxbaum (1969) to the control of the armature current of thyristor fed DC-motors. Buxbaum's analog controller switches from a PI-structure, in continuous current, to an I-structure, in discontinuous current. The realization proposed here is with a microprocessor using fixed-point arithmetic.

The alternative inner robust controller (Unbehauen *et al.*, 1987), which had not yet been so far applied for cascaded speed control of DC motors, will also be used in this investigation.

When the variations in field excitation, moment of inertia, load torque, etc. can be measured or observed, it is also possible to implement parameter scheduling control for the outer speed control loop, as described by Ströle (1967). Another way of adaptation is possible

* Received 2 November 1988; revised 11 October 1989; revised 10 July 1990; received in final form 4 August 1990. The original version of this paper was presented at the IFAC workshop on Robust Adaptive Control which was held in Newcastle, Australia during August 1988. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor G. Verghese under the direction of Editor H. Austin Spang III.

[†] COPPE/UFRJ, P.O. Box 68504, 21945 Rio de Janeiro, Brazil. Author to whom all correspondence should be addressed.

[‡] Ruhr Universität, P.O. Box 102148, 4630 Bochum, Germany.

with an adaptive observer, as proposed by Wehrich and Wohld (1980). The self-tuning approach was studied by Depping and Voits (1982) and Brickwedde (1983). The model reference adaptive approach for the speed control of DC-drives was introduced by Raatz (1970), who applied a control scheme based on the MIT rule. Courtiol and Landau (1975) proposed an analog adaptive system based on hyperstability theory. They used the adaptive model following control (AMFC) approach, where the conditions of perfect model following (Erzberger, 1968) must be satisfied. Recently digital tentative suggestions were made. Lozano and Noriega (1983) applied an adaptive algorithm with forgetting factor, introduced originally by Lozano and Landau (1981), for the digital speed control of a DC servo-motor, without inner current loop. The proposed adaptive controller is not applicable to nonminimum phase plants. Balestrino *et al.* (1983) applied a variant of the adaptive model following control (AMFC) of Courtiol and Landau (1975). Platzer and Kaufman (1984) made simulation studies applying the reference model adaptive control algorithm proposed by Sovel *et al.* (1982).

In the present work the adaptive speed control loop is based on a discrete model reference control strategy introduced by Hahn (1983). Compared with the above mentioned references, it has the following characteristics: (a) stable operation is guaranteed even when the plant is nonminimum phase (Hahn and Unbehauen, 1982), or when the current limits are reached (Hahn, 1985); (b) an inner current control loop is considered; (c) experimental results are presented. The proposed method was already used for the control of a distillation column (Unbehauen and Wiemer, 1985) and of a turbo-generator laboratory system (Hahn *et al.*, 1983). Both implementations were done in mini-computer with standard real-time operating systems. The processes were also sufficiently slow, so that the algorithm calculation time was not such an extremely critical factor as in the present experimental application.

A robust outer speed loop has been implemented in the present investigations for comparison purposes.

Comparisons with a conventional cascade analog control system are also made.

2. BASIC THEORY OF THE IMPLEMENTED ADAPTIVE AND ROBUST CONTROL

The adaptive and robust methods used in this work will be summarized in the following sections.

2.1. The Model Reference Adaptive Controller (MRAC)

2.1.1. *Basic considerations.* The model of the plant with one input $U(z)$, one output $Y(z)$ and unit delay is described by the discrete transfer function

$$G_r(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} z^{-1} = \frac{Y(z)}{U(z)}, \quad (2.1)$$

with the polynomials

$$A(z^{-1}) = 1 + A^*(z^{-1})z^{-1}, \quad (2.2)$$

$$B(z^{-1}) = b_0 + B^*(z^{-1})z^{-1}. \quad (2.3)$$

For simplicity of treatment, only the unit delay case is considered. However, the results can be generalized for plants with arbitrary time delays and multiple inputs and multiple outputs (Hahn, 1983; Unbehauen and Wiemer, 1985). The disturbed plant output is described by (see Fig. 1)

$$A(z^{-1})Y(z) = B(z^{-1})z^{-1}U(z) + V(z), \quad (2.4)$$

$$V(z) = C(z^{-1})\epsilon(z) + Z_d(z), \quad (2.5)$$

where $\epsilon(z)$ is an independent white noise signal and $Z_d(z)$ denotes the unmeasurable deterministic disturbances.

Direct adaptive control schemes like model reference adaptive systems lead to unstable closed loop behaviour with non-minimum phase plants because of the compensation properties of the controller. Therefore, in order to stabilize the nonminimum phase control, the idea of the *correction network* introduced by Hahn and Unbehauen (1982) is applied. The plant output $Y(z)$ is augmented by the signal

$$Y_c(z) = G_c(z^{-1})U(z), \quad (2.6)$$

with

$$G_c(z^{-1}) = \frac{B_c(z^{-1})}{A_c(z^{-1})} z^{-1} = \frac{Y_c(z)}{U(z)}, \quad (2.7)$$

$$A_c(z^{-1}) = 1 + A_c^*(z^{-1})z^{-1}, \quad (2.8)$$

$$B_c(z^{-1}) = b_{c0} + B_c^*(z^{-1})z^{-1}. \quad (2.9)$$

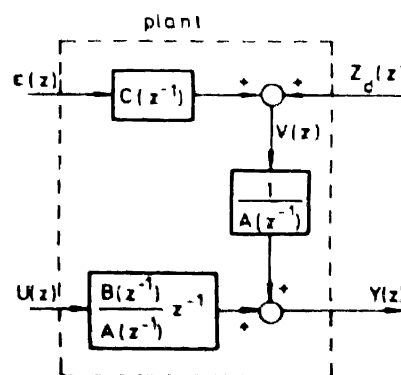


FIG. 1. The basic structure of the plant

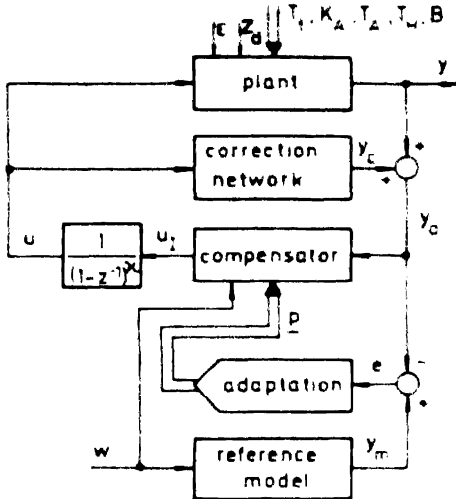


FIG. 2. The basic structure of the MRAC scheme

herein the correction network $G_c(z^{-1})$ should be stable. The augmented plant output is then

$$Y_a(z) = Y(z) + Y_c(z). \quad (2.10)$$

The adaptive controller is now acting on the augmented plant as shown in Fig. 2. Therefore, the adaptive system is closed-loop stable if a stable adaptive algorithm is used and all zeros of the augmented plant lie in the unit circle of the z -plane. These zeros are the roots of the equation

$$G_c(z) + G_p(z) = 0, \quad \text{or} \quad 1 + \frac{G_c(z)}{G_p(z)} = 0. \quad (2.11)$$

If the parameters of the plant are known, the roots of equation (2.11) can easily be investigated with root locus techniques using knowledge of the zeros and poles of the functions $G_c(z)$ and $G_p(z)$. This leads to the following design procedure.

1. Draw the zeros and poles of $G_c(z)$ into the complex z -plane. The zeros of $G_c(z)$ are starting points and the poles are ending points of the branches of the root locus of equation (2.11).
2. Find polynomials $B_c(z)$ and $A_c(z)$, so that all branches of the root locus lie in the unit circle.

Of course, in the case of adaptive control, the plant parameters are essentially not known. Nevertheless, in most cases *a priori* information of the plant will be available and can be used to design the correction network.

In order to reject the influence of deterministic disturbances which have the property

$$\lim_{k \rightarrow \infty} z_d(k) = \text{const.}$$

it is generally convenient to force the controller to have integral action. This can easily be obtained by multiplying equations (2.4) and (2.7) with the factor $(1 - z^{-1})$ on both sides. Therefore, they are modified to

$$A_l(z^{-1})Y(z) = B(z^{-1})z^{-1}U_l(z) + V_l(z) \quad (2.12a)$$

and

$$A_{cl}(z^{-1})Y_l(z) = B_l(z^{-1})z^{-1}U_l(z), \quad (2.12b)$$

with

$$A_l(z^{-1}) = (1 - z^{-1})^k A(z^{-1}) = 1 + A_l^*(z^{-1})z^{-1}, \quad (2.13a)$$

$$U_l(z) = (1 - z^{-1})^k U(z), \quad (2.13b)$$

$$V_l(z) = (1 - z^{-1})^k V(z), \quad (2.13c)$$

$$A_{cl}(z^{-1}) = (1 - z^{-1})^k A_c(z^{-1}) = 1 + A_{cl}^*(z^{-1})z^{-1}, \quad (2.13d)$$

for $k \in \{0, 1\}$. $k=1$ denotes the case with integral action, whereas the integrator is switched off for $k=0$. Note that by this modification and under the assumption of $\epsilon(z) = 0$ and constant deterministic disturbances

$$\lim_{k \rightarrow \infty} v_l(k) = 0 \quad \text{for } k=1. \quad (2.14)$$

This basic structure of the control scheme is shown in Fig. 2.

2.1.2. The control law. The augmented plant output $Y_a(z)$ is compared with the output $Y_m(z)$ of the stable reference model

$$Y_m(z) = G_m(z^{-1})W(z), \quad (2.15)$$

where

$$G_m(z^{-1}) = \frac{B_m(z^{-1})}{A_m(z^{-1})} z^{-1} = \frac{Y_m(z)}{W(z)}, \quad (2.16)$$

and

$$A_m(z^{-1}) = 1 + A_m^*(z^{-1})z^{-1}, \quad (2.17)$$

wherein $W(z)$ is the reference input (set point).

Instead of the model error signal

$$E(z) = Y_m(z) - Y_a(z), \quad (2.18)$$

the filtered model error signal

$$\begin{aligned} E_m(z) &= A_m(z^{-1})E(z) \\ &= A_m(z^{-1})Y_m(z) - A_m(z^{-1})Y_a(z) \\ &\quad - A_m(z^{-1})Y_c(z) \end{aligned} \quad (2.19)$$

is introduced.

Adding the "zero"-term from equations (2.12a,b)

$$\begin{aligned} [A_l(z^{-1})Y(z) - B(z^{-1})z^{-1}U_l(z) - V_l(z)] \\ + [A_{cl}(z^{-1})Y_c(z) - B_l(z^{-1})z^{-1}U_l(z)] = 0 \end{aligned}$$

to the right hand side of equation (2.19) and taking into account from equation (2.16) that

$$A_m(z^{-1})Y_m(z) = B_m(z^{-1})z^{-1}W(z),$$

it follows from equation (2.19) that

$$E_m(z) = z^{-1} [B_m(z^{-1})W(z) + (A_f^*(z^{-1}) - A_m^*(z^{-1}))Y(z) - (B(z^{-1}) + B_f(z^{-1}))U_f(z)] - V_f(z). \quad (2.20)$$

Now, using equations (2.3) and (2.9), the difference equation of the filtered model error is obtained as

$$E_m(z) = z^{-1} [R(z) - (b_o + b_{co})U_f(z) - B^*(z^{-1})z^{-1}U_f(z) + \Delta A^*(z^{-1})Y(z)] - V_f(z), \quad (2.21)$$

with

$$\Delta A^*(z^{-1}) = A_f^*(z^{-1}) - A_m^*(z^{-1}) \quad (2.22)$$

and

$$R(z) = B_m(z^{-1})W(z) - B_f^*(z^{-1})z^{-1}U_f(z) + [A_f^*(z^{-1}) - A_m^*(z^{-1})]Y(z), \quad (2.23)$$

where the signal $R(z)$ is generated only with known parameters and measurable plant input and output signals.

If the control signal is computed by

$$U_f(z) = (b_o + b_{co})^{-1} \times [R(z) - B^*(z^{-1})z^{-1}U_f(z) + \Delta A^*(z^{-1})Y(z)] \quad (2.24)$$

for $V_f(z) = 0$ the augmented plant output $Y_a(z)$ follows the output of the reference model $Y_m(z)$ and $E_m(z) = 0$.

Note that the polynomial $B_f(z^{-1})$ is forced to have a zero at $z = 1$ if integral action is required. Then, under the assumption

$$\lim u(k) = \text{const}, \quad (2.25)$$

the output $y_i(k)$ of the correction network vanishes for $t \rightarrow \infty$. Equation (2.25) can only be satisfied for constant reference signals $W(z)$ and step disturbances $Z_d(z)$. In this case we have

$$\lim y(k) = \lim y_m(k), \quad (2.26)$$

i.e. asymptotic model matching of the plant output.

In equation (2.24) the coefficients of the polynomials $B^*(z^{-1})$ and $\Delta A^*(z^{-1})$ depend on the unknown plant parameters. For the computation of the adaptive control law they have to be replaced by their estimates

$$U_f(z) = (\hat{b}_o + \hat{b}_{co})^{-1} \times [R(z) - \hat{B}^*(z^{-1})z^{-1}U_f(z) + \Delta \hat{A}^*(z^{-1})Y(z)]. \quad (2.27)$$

In equation (2.27) this is indicated by the symbol "...". This equation can easily be solved, provided the term $\hat{b}_o + \hat{b}_{co}$ is not zero. This can

always be achieved by choosing an appropriate b_{co} for the correction network.

With equation (2.13b) the output $U(z)$ of the adaptive controller is

$$U(z) = \frac{1}{(1 - z^{-1})^\kappa} U_f(z), \quad (2.28)$$

with $\kappa \in \{0, 1\}$.

2.1.3. Adaptation of parameters. By substituting $R(z)$ from equation (2.27) in equation (2.21) it follows that

$$E_m(z) = A_m(z^{-1})E(z) = z^{-1}[(\hat{b}_o - b_o)U_f(z) + (\hat{B}^*(z^{-1}) - B^*(z^{-1}))z^{-1}U_f(z) - (\Delta \hat{A}^*(z^{-1}) - \Delta A^*(z^{-1}))Y(z)] - V_f(z). \quad (2.29)$$

Considering now time-varying controller parameters, equation (2.29) can be rewritten in the time domain as

$$e_m(k) = [\mathbf{p} - \hat{\mathbf{p}}(k-1)]^T \mathbf{x}(k-1) - v_f(k), \quad (2.30)$$

where all signals $u_f(k)$ and $y(k)$ and their backward time-shifted values are contained in the signal vector $\mathbf{x}(k)$ and all the corresponding parameters are included into the parameter vectors \mathbf{p} and $\hat{\mathbf{p}}(k)$.

Equation (2.30) describes an estimation problem, which is very well studied at least in the disturbance free and white noise case (e.g. Lozano and Landau, 1981). This estimation problem can be formulated as follows. Find a law for the adaptation of the controller parameter $\hat{\mathbf{p}}$ such that

$$\lim_{k \rightarrow \infty} e_m(k) = 0. \quad (2.31)$$

There are different solutions available for this estimation problem. As the least squares estimation takes more time for computing, in the present approach a *generalized stochastic approximation method* (Hahn, 1983), similar to that proposed by Goodwin *et al.* (1981), has been applied for the estimation of the controller parameters:

$$\hat{\mathbf{p}}(k) = \hat{\mathbf{p}}(k-1) + r(k)G\mathbf{x}(k-1)e_m(k), \quad (2.32)$$

where

$$r(k) = \begin{cases} r^*(k) & \text{if } r^*(k) \geq \frac{1}{\rho \mathbf{x}^T(k-1)G\mathbf{x}(k-1) + \gamma} \\ \text{or} & \end{cases} \quad (2.33)$$

$$\frac{1}{\rho \mathbf{x}^T(k-1)G\mathbf{x}(k-1) + \gamma} \quad \text{otherwise}$$

$$\frac{1}{r^*(k)} = \frac{\lambda_1(k)}{r(k-1)} + \lambda_2(k)\mathbf{x}^T(k-1)G\mathbf{x}(k-1) + \lambda_1(k) \quad (2.34)$$

and the values of

$$G > 0, \quad r(-1) > 0, \quad (2.35a)$$

$$0 \leq \lambda_1(k), \quad \frac{1}{2} < \lambda_2(k), \quad (2.35b)$$

$$0 \leq \lambda_3(k), \quad \rho > \frac{1}{2}, \quad \gamma \geq 0, \quad (2.35c)$$

are freely selectable. It is advisable to choose $\rho \gg \lambda$ and large values of γ . This algorithm, equations (2.32-2.35), contains the "classical" adaptation algorithm with nondecreasing (fixed) gain estimation (e.g. Ionescu and Monopoli, 1977) as a special case setting

$$\lambda_1 = 0, \quad \lambda_2 > 0.5, \quad \lambda_3 = 1. \quad (2.36)$$

The stability of the total algorithm in the disturbance free case, i.e. $v(k) \approx 0$, is guaranteed if $G_c(z^{-1})$ and $G_m(z^{-1})$ are stable and $G_c(z^{-1}) + G_c(z^{-1})$ has no zeros for $|z| \geq 1$. Then all signals are bounded. The algorithm can be extended to the disturbed case by introducing a "dead zone" into the adaptation law (Peterson and Narendra, 1982). The stability proof follows easily using the Liapunov function

$$V(\tilde{\mathbf{p}}(k)) = \tilde{\mathbf{p}}(k)^T G^{-1} \tilde{\mathbf{p}}(k), \quad (2.37)$$

where $\tilde{\mathbf{p}}(k) = \mathbf{p} - \hat{\mathbf{p}}(k)$ is the parameter error.

As the reference model is stable, the convergence of the original error $E(z)$ directly follows. The influence of step disturbances on the estimation procedure is at least asymptotically rejected if an explicit integrator ($\kappa = 1$) is used because then equation (2.14) holds.

2.2. The robust controller

Among several approaches found in the robust control theory, the dominant pole placement concept of Dastych (1983) was chosen for the present investigation. The design goal is to guarantee nearly the same dominant pole configuration of the closed control loop, independently of variations and uncertainties of the plant parameters. The principal characteristics of the method will be summarized in the following sections.

2.2.1. Dominant poles. If the characteristic equation of a control system is written as

$$P(s) = KP_d(s) + \sum_{i=\mu+1}^{\mu+\nu} p_i s^i, \quad K \text{ constant} \quad (2.38)$$

with

$$P_d(s) = \sum_{i=0}^{\mu} p_i s^i \quad \text{and} \quad p_{\mu+1} = 1, \quad (2.39)$$

and if its coefficients are related by

$$p_{\mu} K \gg p_{\mu+1} \quad \text{and} \quad p_i \gg p_{i+1} \quad \text{for} \quad i = \mu + 1, \dots, \mu + \nu - 1 \quad (2.40)$$

then the poles of the polynomial $P_d(s)$ are very close to the dominant poles of $P(s)$. The conditions given by equation (2.40) are satisfied, for example, with the following characteristic polynomial

$$P(s) = P_h(s)P_a(s) = \sum_{i=0}^{\mu+\nu} p_i s^i, \quad (2.41)$$

where

$$P_h(s) = \sum_{i=0}^{\mu} (\delta_i s^i), \quad P_a(s) = (s + \alpha)^{\nu}, \quad (2.42)$$

$$\delta_{\mu} = 1 \quad \text{and} \quad \alpha \rightarrow \infty.$$

In this case, the comparison of equation (2.41) with equation (2.38) gives

$$\lim_{\alpha \rightarrow \infty} \frac{KP_d(s)}{P_a(s)} = P_h(s), \quad (2.43)$$

i.e. $P_h(s)$ is the dominant parcel of $P(s)$ when $\alpha \rightarrow \infty$.

2.2.2. Controller structure and pole placement. The system structure is shown in Fig. 3 where

$$G_s(s) = \frac{D(s)}{C(s)} = \frac{d_0 + d_1 s + \dots + d_m s^m}{c_0 + c_1 s + \dots + s^n}, \quad n > m, \quad (2.44)$$

is the plant transfer function and

$$G_R(s) = \frac{B(s)}{A(s)} = \frac{b_0 + b_1 s + \dots + b_w s^w}{a_0 + a_1 s + \dots + s^q} \quad q \geq w, \quad (2.45)$$

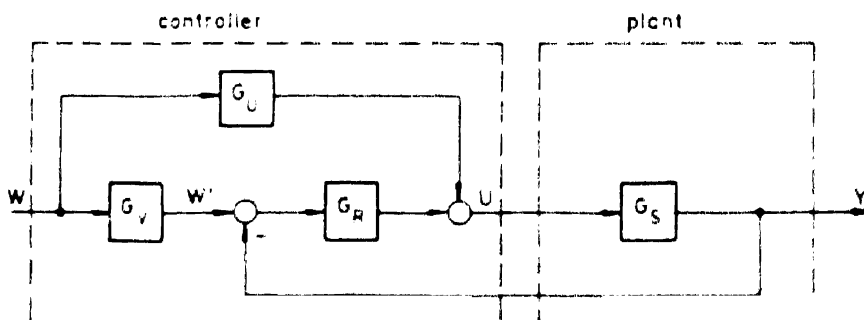


FIG. 3 The robust controller structure

is the controller transfer function.

$$G_V(s) = \frac{M(s)}{N(s)} = \frac{m_0 + m_1 s + \dots + m_r s^r}{n_0 + n_1 s + \dots + s^y}, \quad y \geq x, \quad (2.46)$$

is the transfer function of a pre-filter and

$$G_U(s) = \frac{Z_U(s)}{N_U(s)} \quad (2.47)$$

that of a feedforward path.

Initially G_U and G_V will not be considered. For this condition, the characteristic polynomial of the closed loop is given by

$$P(s) = A(s)C(s) + B(s)D(s) = p_0 + p_1 s + \dots + s^{q+n}. \quad (2.48)$$

For a controller transfer function $G_R(s)$ with integral term, the denominator $A(s)$ in equation (2.45) is given by

$$A(s) = s(a'_0 + a'_1 s + \dots + s^{q-1}), \quad (2.49)$$

and there are $q + w$ unknown controller parameters, which can be established comparing the $q + n$ coefficients of equation (2.48). A necessary condition for the existence of a unique solution for this pole placement problem is that

$$q + w = q + n \rightarrow w = n. \quad (2.50)$$

To maintain the controller order minimal it will be required that

$$q = w. \quad (2.51)$$

The resulting equations relating the coefficients of $A(s)$, $B(s)$, $C(s)$ and $D(s)$ can be written as (Unbehauen, 1982; Dastych, 1983)

$$\mathbf{M}\mathbf{r} = \mathbf{p} - \mathbf{m} \quad (2.52)$$

where

$$\mathbf{M} = \begin{bmatrix} d_0 & 0 & & 0 & 0 \\ & d_1 & d_0 & & c_0 & 0 \\ & & \ddots & \ddots & & \\ \vdots & d_1 & & d_0 & c_1 & c_0 \\ d_m & \vdots & & d_1 & \vdots & c_1 & c_0 \\ & d_m & \vdots & & c_{n-1} & \vdots & c_1 \\ & & \ddots & \ddots & & & \\ & 0 & & d_m & 1 & c_{n-1} & \vdots \\ & & & & & 1 & c_{n-1} \\ & & & & & & 1 \end{bmatrix} \quad (2.53)$$

is a $(2n \times 2n)$ matrix that contains the plant parameters,

$$\mathbf{r}^T = [b_0 b_1 \dots b_n a'_0 a'_1 \dots a'_{n-2}] \quad (2.54)$$

is a vector of controller parameters,

$$\mathbf{p}^T = [p_0 p_1 p_2 \dots p_{2n-1}] \quad (2.55)$$

contains the coefficients of the desired characteristic polynomial, and

$$\mathbf{m}^T = [0 \dots 0 c_0 c_1 \dots c_{n-1}] \quad (2.56)$$

is a vector $(1 \times 2n)$ containing the denominator coefficients of the plant.

2.2.3. Variation of the controller parameters and robust pole-placement. The properties mentioned in section 2.2.1 can be applied to the pole-placement method presented in section 2.2.2. Suppose that the parameters of the plant to be controlled vary within a known region and that a set of nominal parameters, characterized by the index N , is chosen for the controller synthesis. Equation (2.52) gives

$$\mathbf{M}_N \mathbf{r} = \mathbf{p}_N - \mathbf{m}_N. \quad (2.57)$$

The controller parameter vector then is determined by

$$\mathbf{r} = \mathbf{M}_N^{-1}(\mathbf{p}_N - \mathbf{m}_N) \quad (2.58)$$

if \mathbf{M}_N^{-1} exists, i.e. if $D(s)$ and $C(s)$ have no common zeros. The general case is described by equation (2.52). The substitution of equation (2.58) in equation (2.52) leads to

$$\mathbf{p} = \mathbf{M}\mathbf{M}_N^{-1}(\mathbf{p}_N - \mathbf{m}_N) + \mathbf{m}. \quad (2.59)$$

The aim is that \mathbf{p} and \mathbf{p}_N have the same dominant poles. This can be achieved if $P_N(s)$ is chosen as suggested by equation (2.41) and if the elements p_i of the vector \mathbf{p} , given by

$$p_i = K_i p_{iN} + R_i \quad \text{for } i = 0, \dots, 2n, \quad (2.60)$$

satisfy the conditions (2.40). The factor K_i and the remainder terms R_i depend on the plant parameters and on the characteristic equation of the closed loop. These considerations are used to choose the nominal plant parameters and the vector \mathbf{p}_N necessary for the controller synthesis presented in equation (2.58).

The application of this method can be summarized in the following way:

1. Choice of a desired dynamic behaviour through the polynomial $p_d(s)$.
2. Choice of a nominal operating condition, characterized by the index N .
3. Establishment of the controller order using equations (2.50) and (2.51).
4. Choice of α so that equation (2.40) is satisfied.

5. Establishment of the controller by solving equation (2.58).

Finally, the pre-filter $G_V(s)$ and the feed-forward filter $G_U(s)$ can be chosen to improve the system performance at determined operating points. For example, if the desired transfer function of the closed loop is given by

$$\frac{Y(s)}{W(s)} = K_w(s) = \frac{P_{0W}}{P_N(s)}, \quad (2.61)$$

then the pre-filter transfer function

$$G_V(s) = \frac{P_{0N}}{B(s)D_N(s)} \quad (2.62)$$

can be chosen to compensate the zeros of the closed loop, when $G_R(s)$ and $G_S(s)$ are minimum-phase transfer functions. One can also choose

$$G_V(s) = K_w(s) \quad (2.63)$$

and

$$G_U(s) = \frac{K_w(s)}{G_M(s)}, \quad (2.64)$$

where $G_M(s)$ is an approximate model of the plant at a determined operating point and must be chosen so that $G_U(s)$ is realizable. In this case the closed loop transfer function is given by

$$G_W(s) = \frac{Y(s)}{W(s)} = K_w(s)[1 + Q_c(s)], \quad (2.65)$$

where

$$Q_c(s) = \frac{G_S(s) - G_M(s)}{G_M(s)[1 + G_R(s)G_S(s)]}. \quad (2.66)$$

It can be seen that $Q_c(s) = 0$ when the plant transfer function $G_S(s)$ is approximately equal to $G_M(s)$. In this case the system transfer function $G_W(s)$, at the specific operating point, becomes equal to $K_w(s)$.

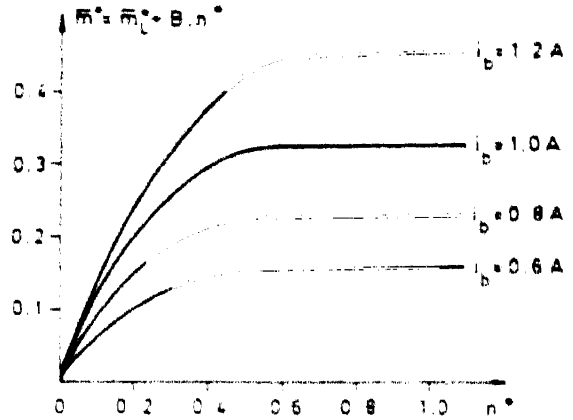


FIG. 5. Dependence of the scaled load torque (m^*) and of the friction coefficient (B) with the excitation current (i_b) of the eddy-current brake upon the scaled speed (n^*)

3. THE PLANT TO BE CONTROLLED

The plant consists of a thyristor fed 1.1 kW separately excited DC-machine, an eddy-current brake, a magnetic clutch, an excitation field supply unit and a tachogenerator. The armature power control rectifier contains two thyristor bridges, each one with four thyristors in fully controlled single-phase configuration. Variations in moment of inertia, field excitation and load are then possible. The single-phase supply makes the discontinuous current domain wide. All these variations make the application of an adaptive or of a robust solution interesting.

The plant was identified by means of classical methods. The mean value model used is explained in detail in Stephan (1987). Figures 4 and 5 present the final model.

4. CONTROLLERS STRUCTURES AND IMPLEMENTATION

The three cascade speed controllers compared in this work are reproduced in Fig. 6. The

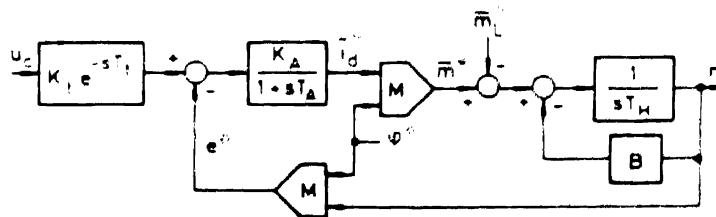


FIG. 4. Scaled plant model (the signals within this nonlinear block diagram are considered to be in the time domain. T_i , K_A , T_A and B are variable parameters).

u_c : rectifier input signal $\begin{cases} u_c = 1 \rightarrow 0^\circ \\ u_c = 0 \rightarrow 90^\circ \\ u_c = -1 \rightarrow 180^\circ \end{cases}$
 e^* : scaled motor back voltage
 i_a^* : scaled armature current (mean value during half period)
 n^* : scaled speed
 m^* : scaled load torque (alterable with the eddy-current brake)
 $(0 < m^* < 0.9)$

$K_i = 2.2$
 $0 < T_i < 10 \text{ ms}$ } rectifier characteristics
 $K_A = 2.0$
 $T_A = 30 \text{ ms}$ } motor/rectifier in
 $0.3 < K_A < 2.0$ } continuous current mode
 $T_A = 0$ } discontinuous current mode
 $0.5 < \varphi^* < 1.0$: scaled field excitation
 $330 \text{ ms} < T_H < 700 \text{ ms}$: electromechanical time constant
 $0 < B < 2$: friction coefficient (alterable with the eddy-current brake)
 Reference values: current 16 A, speed 1900 rpm, motor voltage 130 V, torque 10.4 Nm.

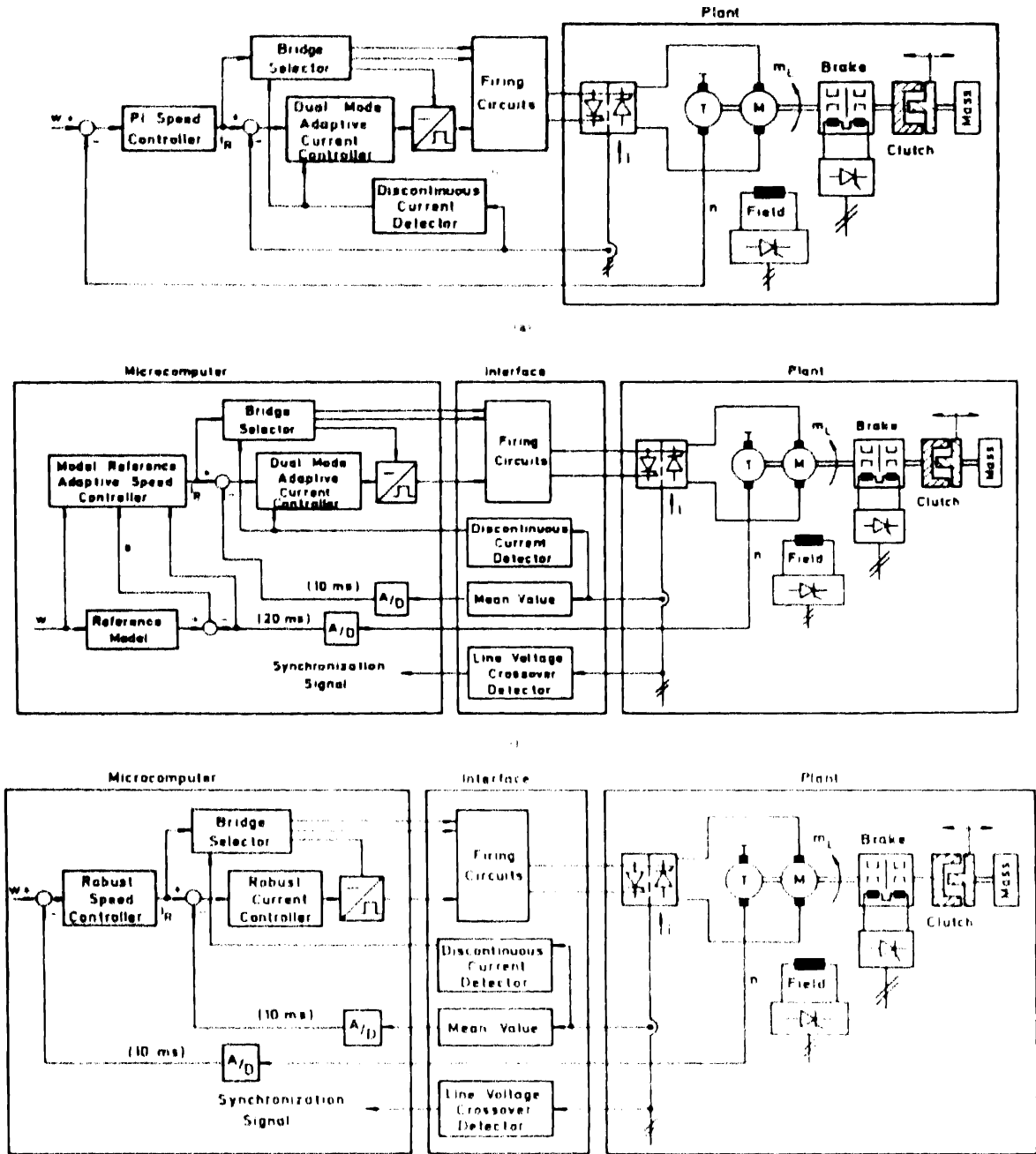


FIG. 6 The cascaded speed control schemes. (a) commercially available analog controller, (b) completely adaptive digital scheme, (c) digital robust scheme

entirely analog controller (Fig. 6a) is a standard commercially available unit (Siemens, Simoreg-6RA21), with dual-mode adaptive inner current loop. The other two schemes were implemented during this investigation on a 16-bit single board computer. It consists of a 5 MHz 8086-CPU, with floating points coprocessor 8087. Moreover, the SBC includes an interrupt controller (8259), a serial (8251) and a parallel (8255) communication interface, a timer (8253), RAM and EPROM's 12 bits A/D and D/A converters.

The dual-mode current adaptive inner controller was programmed in machine language for fix-point arithmetic. It occupies approximately 1 K byte and executes in 0.8 ms. It is important to mention that the inner current loop performs more than a simple control task. Actually, it is also responsible for the commands necessary during a change in the direction of the load current. These functions belong to the "bridge selector" block shown in Fig. 6. The sampling time used for this control loop was 10 ms.

because the single phase thyristor bridge, operating at 50 Hz, allows changes in the firing angle only at intervals of approximately 10 ms.

The model reference adaptive speed controller and the robust controllers were programmed in PASCAL. Assembler subprograms for floating point multiplication, division, subtraction, addition and comparison developed by the authors make the use of the 8087 facilities easy and the computation time nearly minimal. For the robust controller it was possible to use a sampling time of 10 ms and for the adaptive speed controller of 20 ms. These sampling times are respectively 1/10 and 1/5 of the desired time constant of the controlled motor. Practical applications reveal that these are reasonable values for sampling times (Unbehauen, 1985a). Moreover, these times can be directly obtained from the line voltage at 50 Hz.

5 EXPERIMENTAL RESULTS AND COMPARISONS

5.1. The inner current loop

Three different current controllers are tested: the analog dual-mode adaptive, the digital dual-mode adaptive and the digital robust current controllers (Unbehauen *et al.*, 1987). The digital controllers work with a sampling time of 10 ms.

In Fig. 7 the dynamic behaviour of the current

loop for step variations of the reference current are shown for continuous and discontinuous current, as well as for the critical change from discontinuous to continuous current. The analog dual-mode adaptive controller gives in all conditions a rise time of approximately 50 ms and no overshoot. The digital controllers, on the other hand, show a greater variation of the rise time and in two cases a small current overshoot. Nevertheless, the performance of the digital dual-mode adaptive and that of the digital robust controllers are satisfactory for the cascade control. It is interesting to compare these results with the step response without adaptation shown in Fig. 8. The slow transient, with a rise-time greater than 1.2 s, was obtained by switching off the adaptation of the analog or of the digital dual-mode adaptive controllers. This implies that the optimized PI controller for continuous current actuates during the discontinuous current domain too. This slow dynamic behaviour of the inner current loop can make a cascade speed control scheme unstable (Buxbaum and Schierau, 1980).

It is shown, therefore, that the robust controller can substitute with success the dual-mode adaptive current controller and that, in continuous or in discontinuous current conduction mode, the controller inner loop behaves nearly as a first order system with time constant between 15 and 30 ms.

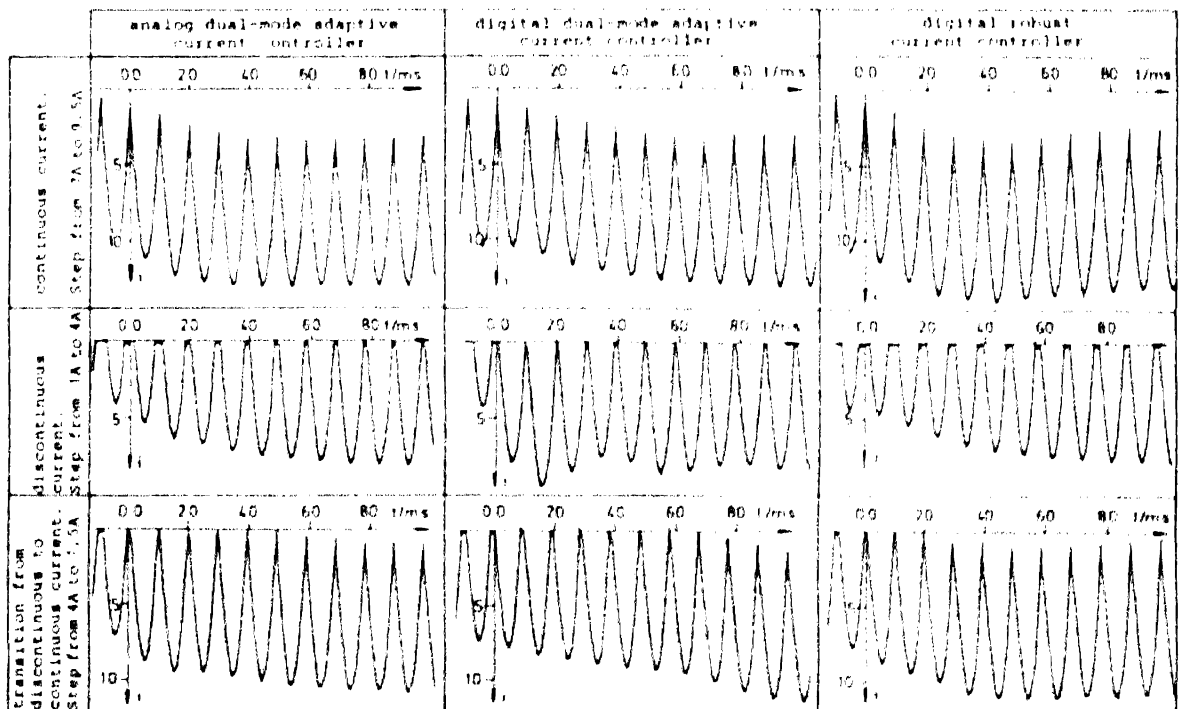


Fig. 7 Step responses for reference current variations under different conditions. i_d current in A

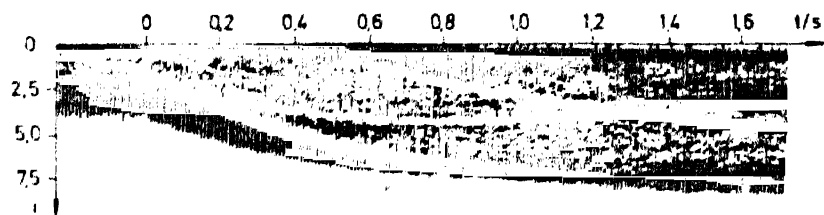


FIG. 8. Current step response in discontinuous current mode (from 1A to 4A); PI-current controller with parameters optimized for continuous current; i = current in A.

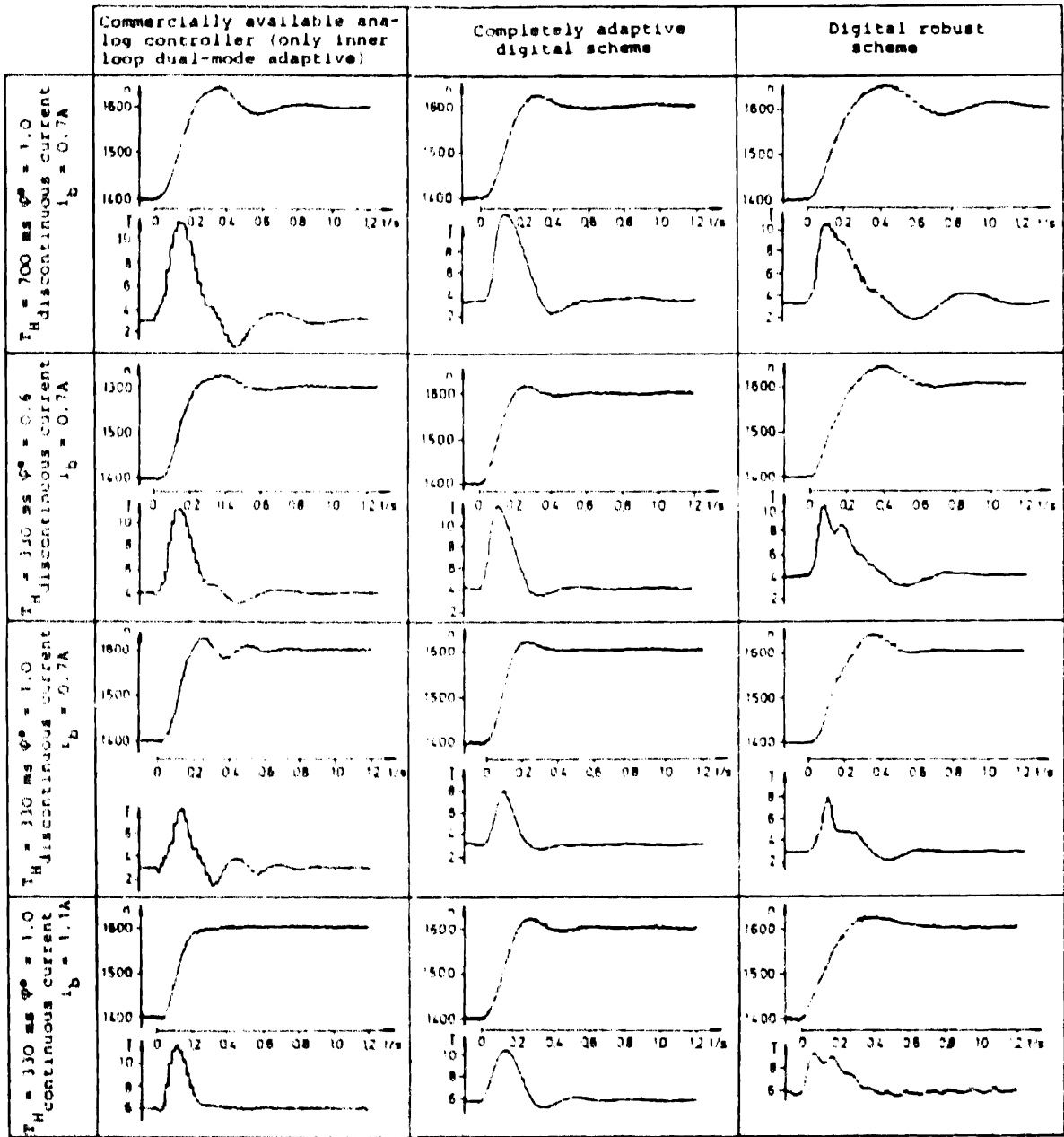


FIG. 9. Step response for change of the reference speed from 1400 to 1600 rpm; n = speed in rpm, i = armature current mean value in A.

5.2. The cascade speed control schemes

5.2.1. Design considerations. The parameters of the commercially available PI-speed controller were optimized using the well known criterion of Kessler (1958). The reference model for the adaptive speed controller was chosen as a first order system with time constant of approximately 0.1 s and gain one. The robust controller was designed to guarantee nearly the same dominant pole configuration of the closed control loop, independently of variations and uncertainties of the plant parameters. The dominant poles were chosen at $-7 \pm j7 \text{ s}^{-1}$, which provides a step response with 5% overshoot and settling time of 0.5 s.

5.2.2. Results. The plant model has already been presented in Figs 4 and 5. The parameters and variables that appear in these figures will be often mentioned during this section. If no comments are made in the text, the results obtained with the adaptive controller are for well adapted parameters to set point variations.

In Fig. 9 the controllers are compared at four different operating points. One can see that the completely adaptive scheme maintains a settling time of 0.5 s and an overshoot smaller than 10%. This is almost the performance given by the parallel model (a first order system with time constant 0.1 s). The differences are due to the correction network, which is necessary in this application because the plant with inner current loop (i.e. the plant to be controlled with the model reference adaptive controller) has non-minimum phase characteristic after discretization with a sampling time of 20 ms (Stephan *et al.*, 1988). On the other hand, the commercially available controller and the robust controller present at some operating conditions an overshoot greater than 10% and a settling time greater than 0.5 s. The advantage of the robust controller, when compared with the commercially available one, is the simpler realization of the inner loop. The latter uses a dual-mode adaptive inner current loop, that changes its structure and parameters at each operating condition, the former has fixed parameters and structure.

The effect of an abrupt change of the moment of inertia can be seen in Fig. 10. This perturbation serves also to show the effect of a pulse of a load torque \bar{m}_l , that occurs when the magnetic clutch is switched on and a mass, initially in repose, is set in motion. After this abrupt variation, the commercially available and the robust controllers show their characteristic responses in the new operating condition. The completely adaptive controller performs well, although without well-adjusted parameters.

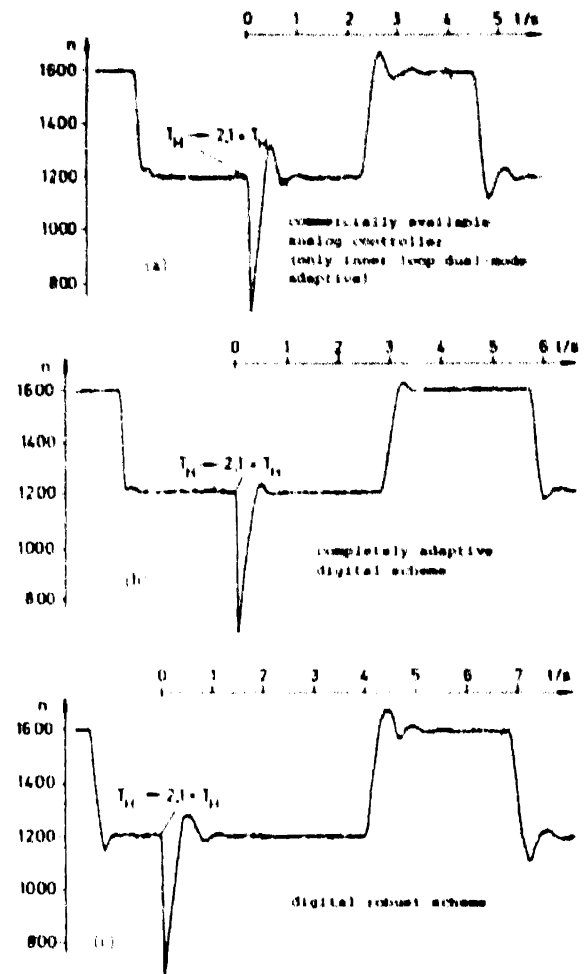


FIG. 10. Perturbation by means of a variation of the moment of inertia and subsequent changes of the reference speed. Conditions: $q^* = 1$, $i_k = 0.8 \text{ A}$, n^* speed in rpm. (a) Commercially available analog controller, (b) completely adaptive digital scheme, (c) digital robust scheme.

A variation of the load torque \bar{m}_l can be obtained altering the excitation current of the eddy-current brake i_k . Figure 11 shows the speed performance in this case. The completely adaptive controller rejects the load perturbation only due to the adaptation. It responds more slowly than the commercially available and robust controllers, that have integral action. As was described in Section 2.2.1, an integral term could also be added to the model reference adaptive controller, but experimental results have shown that such a solution makes the reference current signal (the output of the model reference adaptive controller) wobbly. This fact is expected due to the approximately integral behaviour of the plant at the operating speed of 1500 rpm. As can be seen from Fig. 5, for $n^* = 0.79$, the coefficient B is nearly zero and therefore, according to Fig. 4, the electro-mechanical part of the plant behaves as an

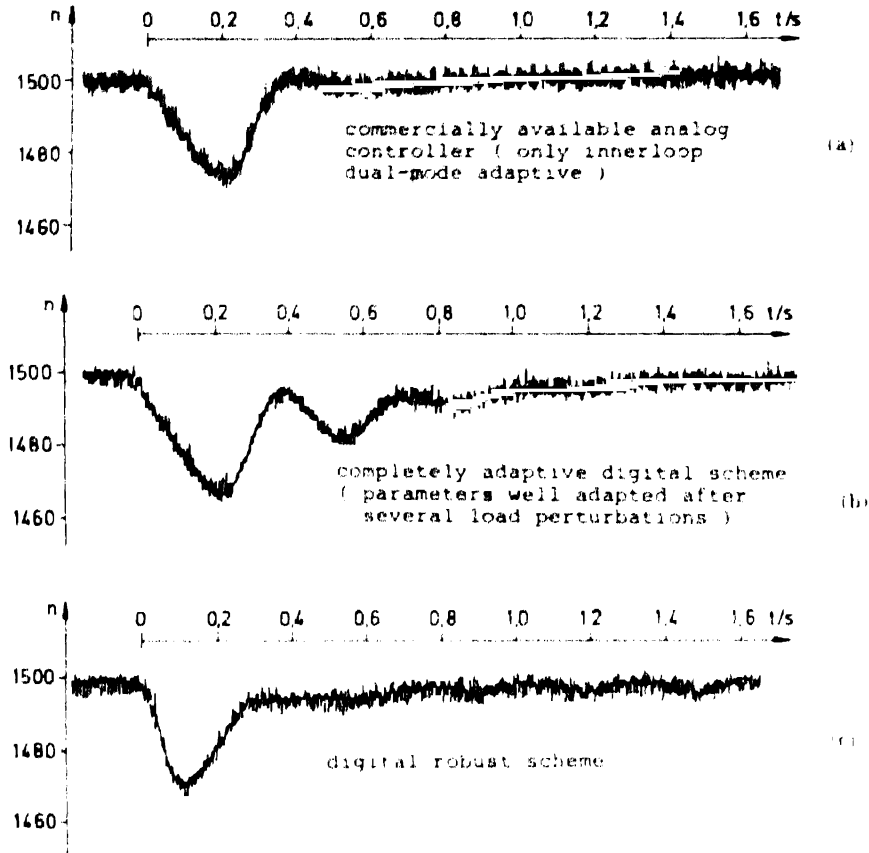


FIG. 11. Perturbation by means of a variation of the current of the eddy-current brake from $i_k = 0.7$ A to $i_k = 1.2$ A. Conditions: $T_H = 330$ ms, $\varphi^* = 1.0$. n $\hat{=}$ speed in rpm. (a) Commercially available analog controller, (b) completely adaptive digital scheme, (c) digital robust scheme

integrator. The speed control loop then, with two integrators, presents a structural instability for the adaptive controller, which explains the wobbly characteristic mentioned above. Moreover, a reference model for set point variations, and not for load perturbation, is

given for the model reference adaptive scheme. This fundamental design consideration handicaps the step response for perturbations and is also responsible for slower parameters adaptation in the case of load perturbation than in the case of setpoint variations.

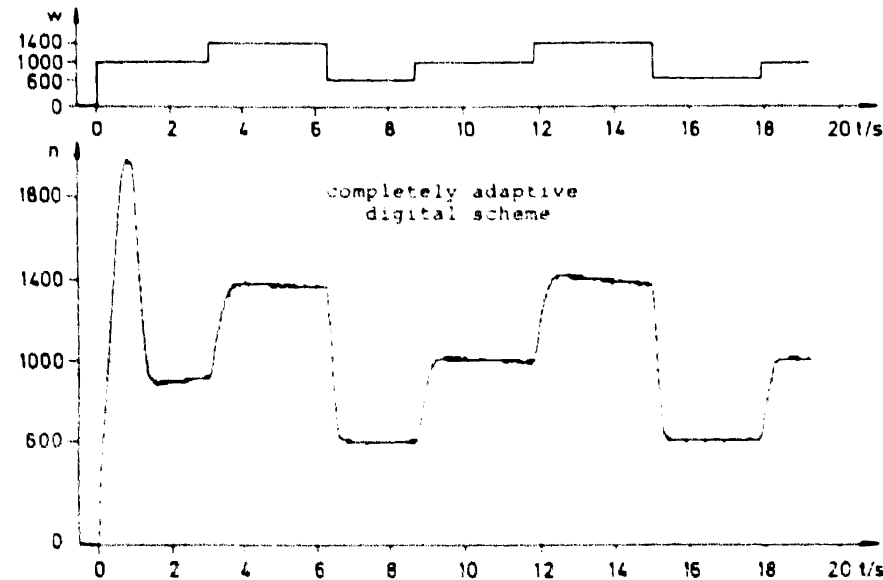


FIG. 12. Adaptation of the controller parameters. Conditions: $T_H = 330$ ms, $\varphi^* = 1.0$, $i_k = 0.8$ A. w $\hat{=}$ reference speed, n $\hat{=}$ speed in rpm

The parameter adaptation for setpoint variations can be judged by the results shown in Fig. 12. All parameters of the model reference adaptive controller are initially set to zero and the adaptive algorithm is then started with a reference speed of 1000 rpm. Then the reference speed is varied between 1400 rpm, 600 rpm and 1000 rpm, furnishing an identification signal. Satisfactory responses are already obtained during the second step. Thus, the controller parameters are adapted.

6. CONCLUSION

A cascade control strategy was chosen for the DC-motor speed control design and three schemes were compared: a commercially available analog scheme, a completely adaptive digital scheme, and a completely robust digital scheme.

The implementation of the digital controllers was made on a standardized 16-bit single board microcomputer.

The obtained results showed that the model reference adaptive control can improve significantly the speed performance for setpoint variations and that the robust current control can substitute with success the usual dual-mode current controller (Stephan, 1985).

REFERENCES

- Balestrino, A., G. Maria and L. Scavico (1983) Adaptive control design in servo-systems. *Preprints 3rd IFAC Symp. on Control in Power Electronics and Electrical Drives*, Lausanne, 125-131.
- Brickwedde, A. (1983) Microprocessor-based Adaptive Control for Electrical Drives. *Preprints 3rd IFAC Symp. on Control in Power Electronics and Electrical Drives*, Lausanne, 119-124.
- Buxbaum, A. (1969) Regelung von Stromrichterantrieben. *Tech. Mitt. AEG*, **59**, 348-352.
- Buxbaum, A. and K. Schierau (1980) *Berechnung von Regelkreisen der Antriebstechnik*. AEG-Telfunken Verlag.
- Courtial, B. and I. Landau (1975) High speed adaptation system for controlled electrical drives. *Automatica*, **11**, 119-127.
- Dastych, J. (1983) Ein Beitrag zum Entwurf von Regelungen für Führungs-Störungsverhalten. Ph.D. Thesis, Ruhr-Universität Bochum, Germany.
- Depping, F. and M. Voits (1982) Microcomputer-based parameter-adaptive speed control with deadbeat response. In W. Leonhard (Ed.) *Microelectronics in Power Electronics and Electrical Drives*, VDE, Berlin, 127-134.
- Erzberger, H. (1968) Analysis and design of model following systems by state space techniques. *Proc. Joint Aut. Control Conf.*, Ann Arbor, 572-581.
- Goodwin, G. C., P. J. Ramadge and P. E. Caines (1981) Discrete time stochastic control. *SIAM J. Control Optimiz.*, **19**, 829-853.
- Hahn, V. (1983) Direkte Adaptive Regelstrategien für die Diskrete Regelung von Mehrgrößensystemen. Ph.D. Thesis, Ruhr-Universität Bochum, Germany.
- Hahn, V. (1985) A direct adaptive controller for non-minimum phase multivariable systems. *Preprints 7th IFAC/IFIP/IMAC Conf. Digital Computer Applications to Process Control*, Vienna, 523-528.
- Hahn, V. and H. Unbehauen (1982) Direct adaptive control of non-minimum phase systems. *Preprints IEEE Conf. on Applications, Adaptive and Multivariable Control*, Hull, U.K. 170-175.
- Hahn, V., H. Unbehauen and V. Nadolph (1983) Model reference adaptive control of a multivariable non-minimum phase plant. *Preprints 3rd Yale Workshop on Applications of Adaptive System Theory*, Yale, 41-46.
- Ionescu, T. and R. V. Monopoli (1977) Discrete model reference adaptive control with an augmented error signal. *Automatica*, **13**, 507-517.
- Kessler, C. (1958) Das Symmetrische Optimum. *Regelungstechnik*, **6**, 395-400, 432-436.
- Kummel, F. (1965) Einfluß der StellgröÙeneigenschaften auf die Dynamik von Drehzahlregelkreisen mit unterlagelter Stromregelung. *Regelungstechnik*, **13**, 227-234.
- Lozano, R. and I. Landau (1981) Redesign of adaptive control schemes. *Int. J. Control*, **33**, 246-268.
- Lozano, R. and A. Noriega (1983) Microcomputer implementation of an adaptive control algorithm. *Preprints IFAC/IFIP Symp. Real Time Digital Control Applications*, Mexico City, 108-113.
- Pelly, B. (1971) *Thyristor Phase-Controlled Converters and Cycloconverters*. Wiley, New York.
- Peterson, B. B. and K. S. Narendra (1982) Bounded error adaptive control. *IEEE Trans. Aut. Control*, **27**, 1161-1168.
- Platzer, D. and H. Kaufman (1984) Model reference adaptive control of thyristor driven DC motor systems. *Preprints 9th World IFAC Conf. Budapest*, **1**, 231-235.
- Raatz, E. (1970) Der Einsatz von Adaptive Drehzahlreglern in der Antriebstechnik. *Tech. Mitt. AEG*, **60**, 375-378.
- Sobel, K., H. Kaufman, I. Mahbus (1982) Implicit adaptive control for MIMO systems. *IEEE Trans. Aerospace*, **18**, 576-590.
- Stephan, R. (1985) Adaptive and robust cascade schemes for thyristor driven DC motor speed control. Ph.D. Thesis, Ruhr-Universität Bochum, Germany.
- Stephan, R. (1987) Global modelling of converter fed DC-motor. *IMACS/IEEE Int. Symp. on Modelling and Simulation of Electrical Machines*, Quebec, 32-39.
- Stephan, R., V. Hahn and H. Unbehauen (1986) Cascade adaptive speed control of a thyristor driven DC motor. *Proc. IEE*, **135**, pt. D, 49-55.
- Ströle, D. (1967) Typische Adaptivsteuerung bei geregelten elektrischen Antrieben. *Regelungstechnik*, **15**, 100-111.
- Unbehauen, H. (1982) *Regelungstechnik I*. Vieweg, Braunschweig, Wiesbaden.
- Unbehauen, H. (1985a) *Regelungstechnik III*. Vieweg, Braunschweig/Wiesbaden.
- Unbehauen, H. (1985b) Theory and application of adaptive control. *Preprints 7th IFAC/IFIP/IMACS Conf. on Digital Computer Applications to Process Control*, Vienna, 3-19.
- Unbehauen, H. and P. Wiemer (1985) Applications of multivariable adaptive control schemes to distillation columns. *Preprints 4th Yale Workshop on Applications of Adaptive System Theory*, Yale, 23-29.
- Unbehauen, H., J. Dastych and R. M. Stephan (1987) Robust current control for a thyristor driven DC-motor. *Preprints 10th World IFAC Conf.*, Munich, **3**, 331-336.
- Wehrich, G. and D. Wöhlid (1980) Adaptive speed control of DC-drives using adaptive observers. *Siemens Forsch. Entwickl. Ber.*, **9**, 283-287.

Quality Control of Binary Distillation Columns via Nonlinear Aggregated Models*

J. LÉVINE† and P. ROUCHON‡

A robust nonlinear control law is designed to reject unknown feed composition disturbances with overall stability. Implementation to real columns as well as comparisons with classical control strategies, show the robustness and flexibility improvements of the method.

Key Words—Control applications, distillation columns, disturbance rejection, model reduction, nonlinear control systems, overall stability, quality control, singular perturbations

1—Using singular perturbation techniques on a physical model of distillation column, an aggregated model is proposed for control purposes. Nonlinear perturbation rejection techniques via static feedback are then applied to this model to reject the feed composition disturbances around every slowly varying reference trajectory; the existence of such a control law and the stability of the overall closed-loop system are proven. Moreover, the obtained control law can be synthesized with measurements commonly available on distillation columns: the product compositions and two inner temperatures. An industrial implementation on a refinery depropanizer of 42 trays is presented. Simulation comparisons with linear and nonlinear geometric control laws, both using the physical model of high dimension, show the robustness and flexibility improvements provided by our method.

Notation

$f = (f_j)_{j=1, \dots, n}$: state function of the physical model (1)
 $\bar{f} = (\bar{f}_j)_{j=1, \dots, r, j_f, s, n}$: state function of the reduced model (18)
 f_j : function defined by (2) for $j = 1, \dots, n$ and corresponding to the physical model, or function defined by (19) for $j = 1, \dots, r, j_f, s, n$ and corresponding to the reduced model
 F : mole flow of the feed
 \bar{F} : steady-state value of F
 H_j : liquid holdup of tray j
 \bar{H}_j : liquid holdup of tray j for the compartment of m trays
 $H = \sum_{j=1}^n \bar{H}_j$
 $\bar{H}_1 = H_1$
 $\bar{H}_r = \sum_{j=1}^r \bar{H}_j$: the liquid holdup of the rectifying compartment
 $\bar{H}_{j_f} = \sum_{j=1}^{j_f-1} \bar{H}_j$: the liquid holdup of the feed compartment.

$H_s = \sum_{j=1}^s \bar{H}_j$: the liquid holdup of the stripping compartment
 $\bar{H}_m = H_m$
 j : tray index
 j_m : index of the aggregation tray relative to the compartment of m trays
 j_f : index of the feed tray
 j_r : defines the rectifying compartment ($2 \leq j_r \leq j_f$)
 j_s : defines the stripping compartment ($j_f \leq j_s \leq n+1$)
 k : the equilibrium function
 $k(\bar{x}_{m+1})$: mole composition of the vapor entering the compartment of m trays
 $k(x_j)$: vapor mole fraction on tray j
 L : reflux flow
 \bar{L} : steady state value of L
 \bar{L} : liquid flow in the compartment of m trays
 m : number of trays of the compartment
 n : number of trays of the column
 r : index of the aggregation tray relative to the rectifying compartment
 s : index of the aggregation tray relative to the stripping compartment
 t : the time
 $T_j = \Theta(x_j)$: temperature of tray j as a function of the liquid composition
 $u = (u_1, u_2)$: the new control vector defined in Theorem 5
 V : reboiler vapor outflow
 \bar{V} : steady state value of V
 \bar{V} : vapor flow in the compartment of m trays
 $x = (x_j)_{j=1, \dots, n}$: state vector of the physical model (1)
 $\bar{x} = (\bar{x}_j)_{j=1, \dots, r, j_f, s, n}$: state vector of the reduced model (18)
 x^s : slow part of the state vector
 x^f : fast part of the state vector
 x_j : liquid mole fraction on tray j for $j = 1, \dots, n$ (a component of the physical state vector), or a component of the reduced state vector for $j = 1, \dots, r, j_f, s, n$ [see (18)]
 \bar{x}_j : steady state value of x_j
 \bar{x}_0 : mole composition of the liquid entering the compartment of m trays
 X^m : function defined by Lemma 1
 $y_1 = x_1$: the quality of the top product
 $y_2 = x_n$: the quality of the bottom product
 \bar{y}_1 : steady state value of y_1
 \bar{y}_2 : steady state value of y_2
 Y^m : function defined by Lemma 1
 z_f : feed composition
 \bar{z}_f : steady state value of z_f

* Received 18 February 1989; revised 8 March 1990; revised 3 August 1990; received in final form 4 September 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Y. Arkum under the direction of Editor H. Austin Spang III.

This work has been partially supported by ELF-FRANCE of the group ELF-AQUITAINE and E.E.C. under contract number EN3E/0134/F.

† École des Mines de Paris, Centre Automatique et Systèmes, 35 rue Saint Honoré, 77305 Fontainebleau, France. Author to whom all correspondence should be addressed.

‡ École des Mines de Paris, Centre Automatique et Systèmes, 60 Bd Saint Michel, 75006 Paris, France.

ϕ_1 : function defined in Theorem 5

ϕ_2 : function defined in Theorem 5

γ : relative to the feed

δ : relative to slow dynamics

ϵ : relative to fast dynamics

β : relative to the compartment of m trays

σ : relative to the steady-state

1 INTRODUCTION

COMPOSITION CONTROL of distillation columns has been studied extensively. The purpose is to maintain the product qualities at their setpoints, even if the feed flowrate and composition vary (generally, the feed composition is not measured). However, very few industrial columns maintain dual composition control. The main reason lies certainly in the difficulties attached to this problem [see for example Fuentes and Luyben (1983)]: strongly nonlinear and interactive system, very sluggish responses, deadtime in the composition measurements and large dimension.

In the literature relative to distillation dynamics and control, two generally separate streams can be identified: several papers emphasize modeling without strong influence on the control design, whereas several other papers concentrate on control without discussing the model.

In the "modeling stream", many physical models are proposed for simulation purposes [see Gallun and Holland (1982) for example]. Since the early study of Rosenbrock (1962), who established most of the qualitative results concerning such models, little theoretical progress has been observed. Other modeling ideas have been developed, for example by España and Landau (1978) and more recently by Benallou *et al.* (1986). They propose compartmental approximation techniques to obtain a simpler model of reduced order.

In the "control stream", many papers have been published and the field can be divided into many parts, in particular: linear predictive control [see for example Morari (1988) and Georgiou *et al.* (1988)]; adaptive control [see for example Agarwal and Seborg (1987)]; linear geometric control [see the pioneering work, using Wonham's (1974) approach, of Takamatsu *et al.* (1979) and more recently Kummel and Andersen (1987)]; and the nonlinear geometric approach [see Gauthier *et al.* (1983) and Lévine and Rouchon (1986)] applying the methods of Hirschorn (1981) and Isidori *et al.* (1981); see also the nonlinear control approach of Alsop and Edgar (1987) based on the approaches of Jakubczyk and Respondek (1980), Hunt *et al.*

(1983) and Krener (1984). Another related approach on extensive variable control can be found in Georgakis (1986).

To summarize, one can observe that firstly, distillation processes are carefully modeled and their dynamical properties are well established; secondly, recent developments of nonlinear control theory are able to provide efficient tools to incorporate the nonlinear aspect of this problem into the control design. In this perspective, our contribution is the following: firstly, we construct a simplified model where the nonlinear and qualitative properties of physical models are preserved (steady-state gains, molar fraction in $[0, 1]$, global asymptotic stability, ...); secondly, we compute a nonlinear control law, rejecting the feed composition perturbations, simple and robust enough to be implemented on a refinery depropanizer.

More precisely, we show that the classical distillation model studied by Rosenbrock (1962) can be approximated, via singular perturbation techniques [see Kokotovic (1984) for example], by a reduced-order model including only the slow transients but having the global asymptotic properties of the original model. This aggregated model proves very useful for control: we apply nonlinear perturbation rejection techniques [see Isidori (1989)] and obtain, around every slowly varying reference trajectory, a feedback law without singularity, producing asymptotically stable closed-loop dynamics, that can be synthesized via output feedback when temperature measurements are available. Though computed from a simplified model, this control law appears to be extremely robust when facing, in an industrial environment, delayed measurements and modeling errors. It is currently being used on a refinery depropanizer, providing, moreover, energy savings and increases in productivity and flexibility.

In the first section, a classical nonlinear physical model of binary distillation columns is recalled. The time-scale aggregation technique is presented and the qualitative properties of the resulting model are analyzed. In the second section, the existence of the control law, rejecting the feed composition perturbations on the aggregated model, and the closed-loop stability are proven. A robust control synthesis via output feedback is proposed and its implementation on a refinery depropanizer is presented. The last section is devoted to simulation comparisons between the obtained control law, the linear geometric one [Takamatsu *et al.* (1979)] and the nonlinear geometric one [inspired by Gauthier *et al.* (1983) and derived from the physical model of this paper].

2. THE CONTROL MODEL

After stating classical modeling assumptions, we recall the associated physical model which is reduced in Section 2.3 by time-scale considerations. For clarity's sake, the reduction is presented firstly on an arbitrary section of trays (compartment) and then extended to the overall column. In the whole section, the major tool for analyzing stability and convergence of our reduction method, is the global stability result of Rosenbrock (1962) that is recalled in Appendix A.

2.1. Modeling assumptions

A general description of distillation processes can be found in the book of Van Winkle (1967). The process studied is a classical binary distillation column [see for example España and Landau (1978)] as displayed in Fig. 1. The following assumptions are introduced:

1. On each tray, liquid and vapor phases are perfectly mixed and are at thermodynamic equilibrium.
2. The liquid molar holdup on each tray is constant. The pressure is constant and uniform. The vapor molar holdup on each tray is negligible.
3. The liquid molar inflow is equal to the liquid molar outflow on each tray. The vapor molar inflow is equal to the vapor molar outflow on

each tray (with the exception of extreme trays 1 and n). The feed is a saturated liquid.

Assumption 1 means that the time constants of the mass transfer between liquid and vapor are much shorter than the resident time of each tray. Assumption 2 states that hydrodynamics, pressure and level dynamics are stable and fast enough to be neglected. This implies perfect level and pressure control. Generally, the tray's geometry, the pressure and level control loops are designed such that 1 and 2 are satisfied for smooth enough inputs F , L and V , respectively the feed flow, the reflux flow and the reboiler vapor outflow. The last assumption, 3, is more restrictive and corresponds to the Lewis hypothesis [see Van Winkle (1967, p. 225)]. The control techniques which are used in this paper can be extended to more general models where energy balances are considered. Such an extension can be done only on a formal ground: the analysis of asymptotic stability of more complex models remains an open problem, even if simulations never display instabilities for reasonable operating conditions. Nevertheless, for several industrial columns—such as depropanizer, deethanizer, ...—assumption 3 provides a good approximation of the inner flow profiles.

Remark 1. For clarity's sake, uniform pressure and saturated liquid feed are considered. However, all the results of this paper remain valid if pressure drops corresponding to a given pressure profile and feed vapor fraction are introduced in the model: Rosenbrock's stability result applies and the closed-loop stability property (Theorem 5) can be extended to this case. It suffices to consider in place of the function k , introduced in Section 2.2, a collection of functions k_j where j denotes the tray index. k_j enjoys the same properties as k s but takes into account the pressure on tray j , $j = 2, \dots, n$. Also, denoting v_{feed} the vapor fraction of the feed which is assumed to be measured or estimated, the liquid and vapor flows in the rectifying and stripping sections must be modified as follows: L and V must be replaced by L and $V + v_{\text{feed}}F$ for the rectifying section; $L + F$ and V must be replaced by $L + (1 - v_{\text{feed}})F$ and V for the stripping section.

We now proceed as follows. Firstly, a physical model is obtained from assumptions 1, 2 and 3 as in España and Landau (1978). For most industrial columns, this model is far too large for control purposes, but it can be used directly for simulation and comparisons. Secondly, this model is reduced by time-scale considerations to produce a satisfactory model for control.



FIG. 1 Schematic diagram of a binary distillation column

2.2. A classical physical model

Under assumptions 1, 2 and 3, the dynamic model of a binary column as displayed in Fig. 1 is derived from the balance equation on each tray for one component

$$\begin{aligned} H_1 \dot{x}_1 &= V k(x_2) - V x_1 \\ H_j \dot{x}_j &= L x_{j-1} + V k(x_{j+1}) - L x_j - V k(x_j), \\ &\quad j = 2, \dots, j_f - 1 \\ H_{j_f} \dot{x}_{j_f} &= L x_{j_f-1} + V k(x_{j_f+1}) - (L + F) x_{j_f} \\ &\quad - V k(x_{j_f}) + F z_f \\ H_j \dot{x}_j &= (L + F) x_{j-1} + V k(x_{j+1}) - (L + F) x_j \\ &\quad - V k(x_j), \quad j = j_f + 1, \dots, n - 1 \\ (H_n \dot{x}_n &= (L + F) x_{n-1} - (L + F - V) x_n - V k(x_n), \end{aligned} \quad (1)$$

where:

j denotes the tray index, $1 \leq j \leq n$;

$j = 1$ corresponds to the reflux drum, $j = j_f$ to the feed tray and $j = n$ to the bottom ($3 \leq j_f \leq n - 1$);

$(H_j)_{(1 \leq j \leq n)}$ are the liquid holdups (constant);

$x = (x_j)_{(1 \leq j \leq n)}$ are the liquid molar fractions;

$k(x_j)$ is the vapor molar fraction; k corresponds to the thermodynamic equilibrium point of the binary mixture [see Prausnitz *et al.* (1980)]. We will call k the equilibrium function;

F and z_f are the feed flow and composition, the perturbation variables; F is measured whereas z_f is not;

L and V are the reflux flow and the reboiler vapor outflow, the control variables.

System (1) is rewritten $\dot{x} = f(x, L, V, F, z_f)$ with

$$f(x, L, V, F, z_f) = (f_j(x, L, V, F, z_f))_{j=1, \dots, n}$$

where

$$\begin{aligned} f_1(x, L, V, F, z_f) &= \frac{V k(x_2) - V x_1}{H_1} \\ &\quad \text{for } j = 2, \dots, j_f - 1: \\ f_j(x, L, V, F, z_f) &= \frac{L x_{j-1} + V k(x_{j+1}) - L x_j - V k(x_j)}{H_j} \\ f_{j_f}(x, L, V, F, z_f) &= \frac{L x_{j_f-1} + V k(x_{j_f+1}) - (L + F) x_{j_f} - V k(x_{j_f}) - F z_f}{H_{j_f}} \\ &\quad \text{for } j = j_f + 1, \dots, n - 1: \\ f_j(x, L, V, F, z_f) &= \frac{(L + F) x_{j-1} + V k(x_{j+1}) - (L + F) x_j - V k(x_j)}{H_j} \\ f_n(x, L, V, F, z_f) &= \frac{(L + F) x_{n-1} - (L + F - V) x_n - V k(x_n)}{H_n} \end{aligned} \quad (2)$$

Notice that f is linear with respect to L , V , F and $F z_f$, and that f is smooth if the equilibrium function k is. We assume once and for all that:

Assumption 1. The inputs L , V , F and z_f are continuous time functions from $[0, +\infty[$ to $[0, +\infty[$ such that for all $t \in [0, +\infty[$, $L(t) < V(t) < L(t) + F(t)$ and $z_f(t) \in]0, 1[$;

Assumption 2. The equilibrium function k and its derivative are continuous functions from $[0, 1]$ to $[0, 1]$, $\frac{dk}{dx}(x) > 0$ for all $x \in [0, 1]$, $k(0) = 0$ and $k(1) = 1$ (see Fig. 2).

Remark 2. Assumption 1 means that the flow of the top and the bottom products is positive. Assumption 2 is satisfied for all binary systems: Malesinski (1965) derives from the second thermodynamic principle that the function k is always an increasing function of x . For more general situations, see the analysis of Kwaalen *et al.* (1985). In practice, the equilibrium function k is obtained by solving the algebraic nonlinear equations of the thermodynamic equilibrium. These equations depend on the particular choice of the thermodynamic model and are generally solved numerically. In the depropanizer application below the model of Soave (1972) is used.

The physical model considered is simpler than the model of Rosenbrock (1962) where vapor holdups and vapor Murphree efficiencies are considered whereas tray hydraulic effects are not taken into account. Our model can be seen as a particular case of Rosenbrock's, which contains twice the number of equations, by neglecting vapor holdups and assuming 100% tray efficiencies. For simplicity reasons, we demonstrate directly, for (1), the Rosenbrock open-loop results (and complete them with a spectrum property) by using the theorem of Appendix A.

Theorem 1. Assuming that Assumptions 1 and 2 hold, we have:

(i) For each initial condition x^0 in the closed

$k(x)$

FIG. 2. The equilibrium function $k(x)$.

subset $[0, 1]^n$, the maximal solution of

$$\dot{x} = f(x, L, V, F, z_f)$$

is defined for every $t \in [0, +\infty[$ and satisfies $x(t) \in [0, 1]^n$ for all $t \in [0, +\infty[$;

(ii) For each L, V, F and z_f , there exists a unique steady-state \bar{x} in $]0, 1[^n$, namely a unique solution of $f(\bar{x}, L, V, F, z_f) = 0$; moreover, if k satisfies $k(x) > x$ for all $x \in]0, 1[$, then \bar{x} satisfies

$$1 > \bar{x}_1 > \bar{x}_2 > \dots > \bar{x}_{n-1} > \bar{x}_n > 0;$$

(iii) If L, V, F and z_f are constant and if $x^0 \in [0, 1]^n$, then the system is Lyapunov-stable [see Arnold (1974, p. 155)] and its solution converges to the unique steady state associated to L, V, F and z_f ; moreover, for every $x \in [0, 1]^n$, the Jacobian matrix $\partial f / \partial x$ has real, distinct and negative eigenvalues.

Proof of (i). It is sufficient to prove that the vector field f is oriented inwards on the boundary ∂D of $D = [0, 1]^n$. ∂D is made of the points $(x_1, \dots, x_n) \in [0, 1]^n$ for which there exists $j \in \{1, \dots, n\}$ such that $x_j = 0$ or $x_j = 1$. At such points, it suffices to prove that $\dot{x}_j \geq 0$ if $x_j = 0$ and $\dot{x}_j \leq 0$ if $x_j = 1$. This directly results from formula (2). \square

Proof of (ii) and (iii). We will prove simultaneously the existence and uniqueness of the steady-state, and the global asymptotic stability by applying Rosenbrock's theorem.

In our case, $p = n$, $\xi_k = H_k x_k$ for $k = 1, \dots, n$, $\Omega = \prod_{k=1}^n [0, H_k]$ and system (1) is rewritten $\dot{\xi} = \phi(\xi)$ where ϕ is continuously differentiable. The dependence of ϕ with respect to L, V, F and z_f is omitted since they are assumed to be constant. The preceding proof of (i) implies that assertion (i) of Rosenbrock's theorem is satisfied. From (1) we see that for $i = 2, \dots, n-1$, $\psi_i = 0$, $\psi_1 = \frac{V-L}{H_1}$ and $\psi_n = \frac{L+F-V}{H_n}$ with ψ_i defined by (30). Assumption 1 implies that assertion (ii) of Rosenbrock's theorem is satisfied.

The Jacobian matrix $\partial \phi / \partial \xi$ is the matrix J of Lemma 2 in Appendix A with $p = n$,

$$\begin{pmatrix} \frac{L}{H_1} & & \frac{L+F}{H_n} \\ & H_{j-1} & \\ & & H_j \end{pmatrix} \begin{matrix} L+F & L+F-V \\ H_n & H_n \end{matrix}$$

and

$$b = \left(\frac{V-L}{H_1}, \frac{V}{H_2} \frac{dk}{dx}(x_2), \dots, \frac{V}{H_n} \frac{dk}{dx}(x_n) \right).$$

Assumptions 1 and 2 imply that the vectors a and b have positive components. Consequently, assertion (iii) of Rosenbrock's theorem is satisfied. Moreover, Lemma 2 implies that the eigenvalues of $\partial \phi / \partial \xi$ are distinct, real and negative. Since $\phi(\xi) = Hf(H^{-1}\xi, L, V, F, z_f)$ with $H = \text{diag}[H_1, \dots, H_n]$, the eigenvalues of $\partial f / \partial x$ have the same property. Assertion (iv) of Rosenbrock's theorem results from the tridiagonal structure of the system.

It remains necessary to prove $\bar{x} \in]0, 1[^n$ and the inequalities of (ii). We know that $\bar{x} \in [0, 1]^n$ satisfies $f(\bar{x}, L, V, F, z_f) = 0$. This is equivalent to

$$Fz_f = (V-L)\bar{x}_1 + (L+F-V)\bar{x}_n \quad (3)$$

$$k(\bar{x}_{j+1}) = \frac{L}{V}\bar{x}_j + \left(1 - \frac{L}{V}\right)\bar{x}_1, \quad j = 1, \dots, j_f - 1 \quad (4)$$

$$\bar{x}_{j+1} = \frac{V}{L+F}k(\bar{x}_j) + \left(1 - \frac{V}{L+F}\right)\bar{x}_n, \quad j = j_f + 1, \dots, n, \quad (5)$$

where (3) is obtained by summing all the equations of (1), (4) corresponds to the sum of the j first equations and (5) to the sum of the $n-j+1$ last equations. If $\bar{x}_1 = 0$, then by induction on j in (4) we have $\bar{x}_j = 0$. Using (5) with $j = j_f + 1$, we have $\bar{x}_n = 0$. This is in contradiction with assumption 1 and (3). Similarly, we obtain that $\bar{x}_1 \neq 1$, $\bar{x}_n \neq 0$ and $\bar{x}_n \neq 1$. Consequently $\bar{x}_1, \bar{x}_n \in]0, 1[$. By induction on j in (4) and (5), we have $\bar{x} \in]0, 1[^n$. If we suppose additionally that $k(x) > x$ for all $x \in]0, 1[$, then relations (4) and (5) give the desired inequalities. \square

2.3. The reduced control model

For industrial columns, the dimension of the above dynamic model is generally large (for a refinery depropanizer $n = 40$). It can be reduced by time-scale considerations. The standard form of a two-scale system is [see Kokotovic (1984) or Marino and Kokotovic (1988) for example]

$$\begin{cases} \dot{x}^S = f^S(x^S, x^F, u, w, \varepsilon) \\ \varepsilon \dot{x}^F = f^F(x^S, x^F, u, w, \varepsilon) \end{cases} \quad (6)$$

where (x^S, x^F) is the state vector, the superscript S (resp. F) standing for slow (resp. fast). u is the control vector, w the perturbation vector and ε a small positive scalar. Under suitable assumptions, such a system can be reduced to its slow dynamics [by application of Tikhonov's theorem (Tikhonov *et al.*, 1980) recalled in Appendix B)]

$$\begin{cases} \dot{x}^S = f^S(x^S, x^F, u, w, 0) \\ 0 = f^F(x^S, x^F, u, w, 0) \end{cases} \quad (7)$$

corresponding to $\varepsilon = 0$.

The model (1) is not in standard form (6). Nevertheless for physical reasons (the behavior of each tray is similar to that of any other, the resident time in one tray is much shorter than the resident in a "large" section of trays), we propose a choice of ϵ and a diffeomorphic change of variables to express system (1) in standard form. More precisely, this can be done by splitting the column into a given number of sections of consecutive trays (called compartments), and by aggregation of each section separately. We now present the aggregation method on a given section of trays which gives an alternative model to the compartmental models of Benallou *et al.* (1986) or of España and Landau (1978).

2.3.1. The reduced model of a section of m trays.

Preliminary results. Consider the section of m trays displayed in Fig. 3. Its dynamic model is

$$\begin{aligned} \dot{\tilde{H}}_1 \tilde{x}_1 &= \tilde{L} \tilde{x}_0 + \tilde{V} k(\tilde{x}_2) - \tilde{L} \tilde{x}_1 - \tilde{V} k(\tilde{x}_1) \\ \dot{\tilde{H}}_2 \tilde{x}_2 &= \tilde{L} \tilde{x}_1 + \tilde{V} k(\tilde{x}_3) - \tilde{L} \tilde{x}_2 - \tilde{V} k(\tilde{x}_2) \\ &\vdots \\ \dot{\tilde{H}}_{m-1} \tilde{x}_{m-1} &= \tilde{L} \tilde{x}_{m-2} + \tilde{V} k(\tilde{x}_m) - \tilde{L} \tilde{x}_{m-1} \\ &\quad - \tilde{V} k(\tilde{x}_{m-1}) \\ \dot{\tilde{H}}_m \tilde{x}_m &= \tilde{L} \tilde{x}_{m-1} + \tilde{V} k(\tilde{x}_{m+1}) - \tilde{L} \tilde{x}_m - \tilde{V} k(\tilde{x}_m), \end{aligned} \tag{8}$$

where:

- $(\tilde{H}_i)_{(1 \leq i \leq m)}$ are the liquid holdups;
- $(\tilde{x}_i)_{(1 \leq i \leq m)}$ are the liquid compositions;
- \tilde{L} and \tilde{V} are the liquid and vapor flows entering the section;
- \tilde{x}_0 and $k(\tilde{x}_{m+1})$ are the compositions of the liquid and of the vapor entering the section (k is the previously defined equilibrium function).

The above system is denoted

$$\dot{\tilde{x}} = \tilde{f}(\tilde{x}, \tilde{L}, \tilde{V}, \tilde{x}_0, \tilde{x}_{m+1}),$$

where $\tilde{x} = (\tilde{x}_i)_{(1 \leq i \leq m)}$ and \tilde{f} denotes the right

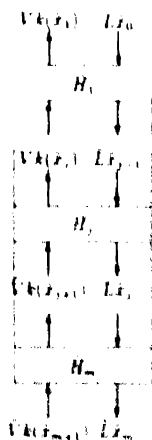


FIG. 3. A compartment of m trays

hand side of (8). Similarly to the overall column, we have the following open-loop behavior.

Theorem 2. Assume assumption 2 and that the inputs \tilde{L} , \tilde{V} , \tilde{x}_0 and \tilde{x}_{m+1} are continuous-time functions such that for all $t \in [0, +\infty[$, $\tilde{L}(t)$, $\tilde{V}(t) \in]0, +\infty[$, $\tilde{x}_0(t) \in [0, 1]$ and $\tilde{x}_{m+1}(t) \in [0, 1]$. Then the following assertions hold true:

- (i) For each initial condition \tilde{x}^0 in $[0, 1]^m$, the maximal solution of (8) is defined on $[0, +\infty[$ and for all $t \in [0, +\infty[$, $\tilde{x}(t)$ remains in $[0, 1]^m$;
- (ii) For each \tilde{L} , \tilde{V} , \tilde{x}_0 and \tilde{x}_{m+1} there exists a unique $\hat{x} = (\hat{x}_i)_{(1 \leq i \leq m)}$ in $[0, 1]^m$ such that

$$\tilde{f}(\hat{x}, \tilde{L}, \tilde{V}, \tilde{x}_0, \tilde{x}_{m+1}) = 0;$$

- (iii) If the functions \tilde{L} , \tilde{V} , \tilde{x}_0 and \tilde{x}_{m+1} are constant and if the initial condition \tilde{x}^0 lies in $[0, 1]^m$, then the system is Lyapunov-stable and

$$\lim_{t \rightarrow +\infty} \tilde{x}(t) = \hat{x}$$

where $\tilde{f}(\hat{x}, \tilde{L}, \tilde{V}, \tilde{x}_0, \tilde{x}_{m+1}) = 0$ defines the steady-state \hat{x} .

The proof is a straightforward adaptation of the proof of Theorem 1 and left to the reader. \square

In the sequel we will use the following lemma relative to the steady-state \hat{x} of (8).

Lemma 1. Assume assumption 2 and that $\tilde{L} > 0$, $\tilde{V} > 0$ and $0 \leq \tilde{x}_0, \tilde{x}_{m+1} \leq 1$. Consider the unique steady-state (Theorem 2) $\hat{x} \in [0, 1]^m$ of (8). Then $k(\hat{x}_1)$ and \hat{x}_m are continuously differentiable functions of \tilde{L}/\tilde{V} , $\tilde{x}_0, \tilde{x}_{m+1}$ denoted respectively $Y^m(\tilde{L}/\tilde{V}, \tilde{x}_0, \tilde{x}_{m+1})$ and $X^m(\tilde{L}/\tilde{V}, \tilde{x}_0, \tilde{x}_{m+1})$. They are related by the equation

$$\begin{aligned} \tilde{L} X^m(\tilde{L}/\tilde{V}, \tilde{x}_0, \tilde{x}_{m+1}) + \tilde{V} Y^m(\tilde{L}/\tilde{V}, \tilde{x}_0, \tilde{x}_{m+1}) \\ = \tilde{L} \tilde{x}_0 + \tilde{V} k(\tilde{x}_{m+1}) \end{aligned} \tag{9}$$

and satisfy

$$\begin{aligned} 0 \leq Y^m \leq 1 & \quad 0 \leq X^m \leq 1 \\ 0 < \frac{\partial Y^m}{\partial \tilde{x}_0} < \frac{\tilde{L}}{\tilde{V}} & \quad 0 < \frac{\partial X^m}{\partial \tilde{x}_0} < 1 \\ 0 < \frac{\partial Y^m}{\partial \tilde{x}_{m+1}} < \frac{dk}{dx}(\tilde{x}_{m+1}) & \quad 0 < \frac{\partial X^m}{\partial \tilde{x}_{m+1}} \\ & \quad < \frac{\tilde{V}}{\tilde{L}} \frac{dk}{dx}(\tilde{x}_{m+1}). \end{aligned} \tag{10}$$

Moreover, if \hat{x} satisfies $\hat{x}_0 > \hat{x}_1 > \dots > \hat{x}_m > \hat{x}_{m+1}$, then

$$\frac{\partial Y^m}{\partial \tilde{L}/\tilde{V}} > 0 \quad \text{and} \quad \frac{\partial X^m}{\partial \tilde{L}/\tilde{V}} > 0. \tag{11}$$

Proof. Notice that (9) results from the sum of all the equations of (8) at the steady state. It remains to prove (10) and (11). The proof is similar for X^m and Y^m . Details are only given for Y^m . We will proceed by induction on m , the number of trays, to prove (10) and (11) for Y^m .

Let us prove that the result is true for $m = 1$ and Y^1 . \hat{x}_1 is given by (8):

$$\frac{\bar{L}}{\bar{V}} \hat{x}_1 + k(\hat{x}_1) = \frac{\bar{L}}{\bar{V}} \hat{x}_0 + k(\hat{x}_2);$$

since k is a continuously differentiable bijection from $[0, 1]$ to $[0, 1]$, \hat{x}_1 exists, is unique and belongs to $[0, 1]$; $k(\hat{x}_1)$ is a regular function Y^1 of \bar{L}/\bar{V} , \hat{x}_0 and \hat{x}_2 . The inequalities relative to Y^1 result from the derivation of the previous equation.

Assume that the result holds for $m - 1 \geq 1$ and Y^{m-1} . Consider $\hat{x} = (\hat{x}_i)_{i=1, \dots, m}$ the steady-state of (8). Then $(\hat{x}_2, \dots, \hat{x}_m)$ is the steady-state of the section made of trays 2 to m corresponding to the $m - 1$ last equation of (8). By the induction assumption, we have

$$k(\hat{x}_2) = Y^{m-1}(\bar{L}/\bar{V}, \hat{x}_1, \hat{x}_{m+1}),$$

where Y^{m-1} is continuously differentiable and satisfies

$$\begin{aligned} 0 &\leq Y^{m-1} \leq 1 \\ 0 &\leq \frac{\partial Y^{m-1}}{\partial \hat{x}_1} \leq \frac{\bar{L}}{\bar{V}} \\ 0 &\leq \frac{\partial Y^{m-1}}{\partial \hat{x}_{m+1}} \leq \frac{dk}{dx}(\hat{x}_{m+1}). \end{aligned}$$

Moreover, if $\hat{x}_1 > \hat{x}_2 > \dots > \hat{x}_i > \hat{x}_{m+1}$, we have

$$\frac{\partial Y^{m-1}}{\partial \bar{L}/\bar{V}} > 0.$$

\hat{x}_1 is then given by

$$\frac{\bar{L}}{\bar{V}} \hat{x}_1 + k(\hat{x}_1) - Y^{m-1}(\bar{L}/\bar{V}, \hat{x}_1, \hat{x}_{m+1}) - \frac{\bar{L}}{\bar{V}} \hat{x}_0 = 0, \quad (12)$$

the equation of (8) corresponding to the tray 1. The left hand side of the above equation is an increasing continuously differentiable function of \hat{x}_1 (use inequality concerning the derivatives of Y^{m-1} with respect to \hat{x}_1), nonpositive for $\hat{x}_1 = 0$ and non-negative for $\hat{x}_1 = 1$. Consequently, this function has a unique zero $\hat{x}_1 \in [0, 1]$. $k(\hat{x}_1)$ is a continuously differentiable function, Y^{m-1} depending on \bar{L}/\bar{V} , \hat{x}_0 and \hat{x}_{m+1} with values in $[0, 1]$. The derivatives with respect to \hat{x}_0 , \hat{x}_{m+1} and \bar{L}/\bar{V} of (12) immediately give the desired inequalities concerning the derivatives of Y^m . \square

Time-scale reduction. We suppose that the trays

are comparable and that $1 \ll m$. Denote $\bar{H} = \sum_1^m \bar{H}_i$ the section holdup. Consider the tray numbered j_a , $j_a \in \{1, \dots, m\}$, called the aggregation tray. For each $j \neq j_a$, we set $\bar{H}_j = \epsilon \alpha_j \bar{H}$ with $0 < \epsilon \ll 1$ and $\alpha_j \approx 1$. We have $\bar{H}_{j_a} = \bar{H} \left(1 - \sum_{j \neq j_a} \epsilon \alpha_j\right)$. Consider the following change of state coordinate associated to j_a

$$\begin{pmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_{j_a-1} \\ \hat{x}_{j_a} \\ \vdots \\ \hat{x}_{j_a+1} \\ \vdots \\ \hat{x}_m \end{pmatrix} \rightarrow \begin{pmatrix} \hat{x}_1^t = \hat{x}_1 \\ \vdots \\ \hat{x}_{j_a-1}^t = \hat{x}_{j_a-1} \\ \hat{x}^\lambda = \left(\sum_1^m \bar{H}_i \hat{x}_i \right) / \bar{H} \\ \hat{x}_{j_a+1}^t = \hat{x}_{j_a+1} \\ \vdots \\ \hat{x}_m^t = \hat{x}_m \end{pmatrix}, \quad (13)$$

where the composition of each tray $i \neq j_a$ remains unchanged and where the composition on the tray j_a is replaced by the weighted sum of the compositions on the m trays of the section. The state equations become

$$\begin{cases} \epsilon \alpha_1 \bar{H} \dot{\hat{x}}_1^t = \bar{L} \hat{x}_0 + \bar{V} k(\hat{x}_2^t) - \bar{L} \hat{x}_1^t - \bar{V} k(\hat{x}_1^t) \\ \vdots \\ \epsilon \alpha_{j_a-1} \bar{H} \dot{\hat{x}}_{j_a-1}^t = \bar{L} \hat{x}_{j_a-2}^t + \bar{V} k \left(\frac{\hat{x}^\lambda - \epsilon \sum_{j \neq j_a} \alpha_j \hat{x}_j^t}{1 - \epsilon \sum_{j \neq j_a} \alpha_j} \right) - \bar{L} \hat{x}_{j_a-1}^t - \bar{V} k(\hat{x}_{j_a-1}^t) \\ \bar{H} \dot{\hat{x}}^\lambda = \bar{L} \hat{x}_0 + \bar{V} k(\hat{x}_{m+1}) - \bar{L} \hat{x}_m^t - \bar{V} k(\hat{x}_1^t) \\ \epsilon \alpha_{j_a+1} \bar{H} \dot{\hat{x}}_{j_a+1}^t = \bar{L} \left(\frac{\hat{x}^\lambda - \epsilon \sum_{j \neq j_a} \alpha_j \hat{x}_j^t}{1 - \epsilon \sum_{j \neq j_a} \alpha_j} \right) + \bar{V} k(\hat{x}_{j_a+2}^t) - \bar{L} \hat{x}_{j_a+1}^t - \bar{V} k(\hat{x}_{j_a+1}^t) \\ \vdots \\ \epsilon \alpha_m \bar{H} \dot{\hat{x}}_m^t = \bar{L} \hat{x}_{m-1}^t + \bar{V} k(\hat{x}_{m+1}) - \bar{L} \hat{x}_m^t - \bar{V} k(\hat{x}_m^t) \end{cases} \quad (14)$$

and are in the standard form (6). The slow and fast subsystems are thus described respectively by

$$\begin{cases} \bar{H} \dot{\hat{x}}^\lambda = \bar{L} \hat{x}_0 + \bar{V} k(\hat{x}_{m+1}) - \bar{L} \hat{x}_m^t - \bar{V} k(\hat{x}_1^t) \\ 0 = \bar{L} \hat{x}_0 + \bar{V} k(\hat{x}_2^t) - \bar{L} \hat{x}_1^t - \bar{V} k(\hat{x}_1^t) \\ \vdots \\ 0 = \bar{L} \hat{x}_{j_a-2}^t + \bar{V} k(\hat{x}^\lambda) - \bar{L} \hat{x}_{j_a-1}^t - \bar{V} k(\hat{x}_{j_a-1}^t) \\ 0 = \bar{L} \hat{x}^\lambda + \bar{V} k(\hat{x}_{j_a+2}^t) - \bar{L} \hat{x}_{j_a+1}^t - \bar{V} k(\hat{x}_{j_a+1}^t) \\ \vdots \\ 0 = \bar{L} \hat{x}_{m-1}^t + \bar{V} k(\hat{x}_{m+1}) - \bar{L} \hat{x}_m^t - \bar{V} k(\hat{x}_m^t) \end{cases} \quad (15)$$

and

$$\left\{ \begin{array}{l} \alpha_1 \tilde{H} \frac{d\tilde{x}_1^t}{d\tau} = \tilde{L}\tilde{x}_0 + \tilde{V}k(\tilde{x}_2^t) - \tilde{L}\tilde{x}_1^t - \tilde{V}k(\tilde{x}_1^t) \\ \vdots \\ \alpha_{j_a-1} \tilde{H} \frac{d\tilde{x}_{j_a-1}^t}{d\tau} = \tilde{L}\tilde{x}_{j_a-2}^t + \tilde{V}k(\tilde{x}^t) \\ \quad - \tilde{L}\tilde{x}_{j_a-1}^t - \tilde{V}k(\tilde{x}_{j_a-1}^t) \\ \alpha_{j_a+1} \tilde{H} \frac{d\tilde{x}_{j_a+1}^t}{d\tau} = \tilde{L}\tilde{x}^t + \tilde{V}k(\tilde{x}_{j_a+2}^t) \\ \quad - \tilde{L}\tilde{x}_{j_a+1}^t - \tilde{V}k(\tilde{x}_{j_a+1}^t) \\ \vdots \\ \alpha_m \tilde{H} \frac{d\tilde{x}_m^t}{d\tau} = \tilde{L}\tilde{x}_{m-1}^t + \tilde{V}k(\tilde{x}_{m+1}^t) \\ \quad - \tilde{L}\tilde{x}_m^t - \tilde{V}k(\tilde{x}_m^t), \end{array} \right. \quad (16)$$

with $\tau = t/\varepsilon$ in (16). We denote $\tilde{x}^t = (\tilde{x}_i^t)_{i \in I_a}$.

Theorem 3. We assume that the equilibrium function satisfies Assumption 2 and that \tilde{L} , \tilde{V} , \tilde{x}_0 and \tilde{x}_{m+1} are continuous time functions such that for all $t \in [0, +\infty[$, $\tilde{L}(t)$, $\tilde{V}(t) \in [0, +\infty[$, $\tilde{x}_0(t) \in [0, 1]$ and $\tilde{x}_{m+1}(t) \in [0, 1]$ for all $t \geq 0$. If the initial conditions of (14), $(x^{s,0}, x^{f,0})$, and of (15), $x^{s,0}$, have their components in $[0, 1]$, then (14) and (15) admit continuous solutions on $[0, +\infty[$, denoted respectively $(\tilde{x}^s(t, \varepsilon), \tilde{x}^f(t, \varepsilon))$ and $(\tilde{x}_0^s(t), \tilde{x}_0^f(t))$, satisfying

$$\lim_{\varepsilon \rightarrow 0} \begin{pmatrix} \tilde{x}^s(t, \varepsilon) \\ \tilde{x}^f(t, \varepsilon) \end{pmatrix} = \begin{pmatrix} \tilde{x}_0^s(t) \\ \tilde{x}_0^f(t) \end{pmatrix},$$

uniformly on every interval of the form $[\alpha, T]$ with $0 < \alpha < T$.

Proof. Let us verify that all the assertions of the Tikhonov's theorem (Tikhonov *et al.*, 1980) (see Appendix B) are satisfied:

- Existence of the solution of the perturbed system (14): Theorem 2 proves that (14) admits a solution $(\tilde{x}^s(t, \varepsilon), \tilde{x}^f(t, \varepsilon))$ on $[0, +\infty[$;
- Existence and stability of the fast subsystem (16): in (16) \tilde{L} , \tilde{V} , \tilde{x}_0 , \tilde{x}_{m+1} and \tilde{x}^s are constant parameters; consequently, (16) is made of two decoupled section of trays, the first one corresponding to the trays 1 to $j_a - 1$, the second one to the trays $j_a + 1$ to m ; Theorem 2 implies that if $\tilde{x}^s \in [0, 1]$, system (16) admits a unique steady state, and if the initial condition lies in $[0, 1]^{m-1}$, then (16) is globally asymptotically stable; moreover, its Jacobian matrix has distinct, real and negative eigenvalues;
- Existence of the solution of the slow subsystem (15): if $\tilde{x}_0^s \in [0, 1]$, the algebraic equations of (15) has a unique solution, \tilde{x}_0^f ,

corresponding to the steady state of (16); using Lemma 1, (15) becomes

$$\begin{aligned} \tilde{H}\tilde{x}_0^s &= \tilde{L}\tilde{x}_0 + \tilde{V}k(\tilde{x}_{m+1}) - \tilde{V}Y^{n-1}(\tilde{L}/\tilde{V}, \tilde{x}_0, \tilde{x}_0^s) \\ &\quad - \tilde{L}X^{n-1}(\tilde{L}/\tilde{V}, \tilde{x}_0^s, \tilde{x}_{m+1}); \end{aligned}$$

the proof of the existence of the solution for $t \in [0, +\infty[$ and remaining in $[0, 1]$ is similar to the proof of (i) in Theorem 1 [use (10) of Lemma 1] and is left to the reader. \square

Remark 3. Expressed in a less mathematical form, the result of Theorem 3 becomes very simple. The dynamics of the whole section can be approximated by the dynamics of a section where the holdup profile is modified as follows: the trays $j \neq j_a$ have no holdup ($0 \rightarrow \tilde{H}_j$), the tray j_a has the section holdup $\left(\sum_{i=1}^m \tilde{H}_i \rightarrow \tilde{H}_{j_a}\right)$. Notice

that the global holdup remains unchanged.

2.3.2. The reduced model of the column. Consider now the overall column. The choice of the compartments (sections of consecutive trays) has to take into account several considerations concerning the holdups. For most columns, the holdup profile is as follows:

- The holdups on external trays 1 and n (reflux drum and bottom) are much more important than the holdups on any other tray (trays 2, ..., $n-1$);
- The holdups on external trays are comparable to the global holdup of all other trays;
- The holdups on trays 2 to $n-1$ are comparable.

The aggregated model should simultaneously have a small dimension and represent correctly the column dynamics.

We can consider that the two external trays have their own slow dynamics. For the trays 2 to $n-1$, the number of compartments constitutes a degree of freedom. In the case of the depropanizer described in the discussion below, open-loop trajectory comparisons between the physical model and different aggregated models (aggregated models of orders 3, 4 and 5, corresponding to aggregations of trays 2 to $n-1$ in respectively 1, 2 and 3 compartments), are displayed in Fig. 4. They correspond to variations of the feed composition slightly more severe than what is usually observed in practice. They show that a good tradeoff between accuracy and dimension can be obtained with an aggregated model of order 5.

In the sequel, we consider a 5-compartment aggregated model as displayed in Fig. 5:

- The two external trays remain unchanged;
- The other trays are decomposed into 3 similar

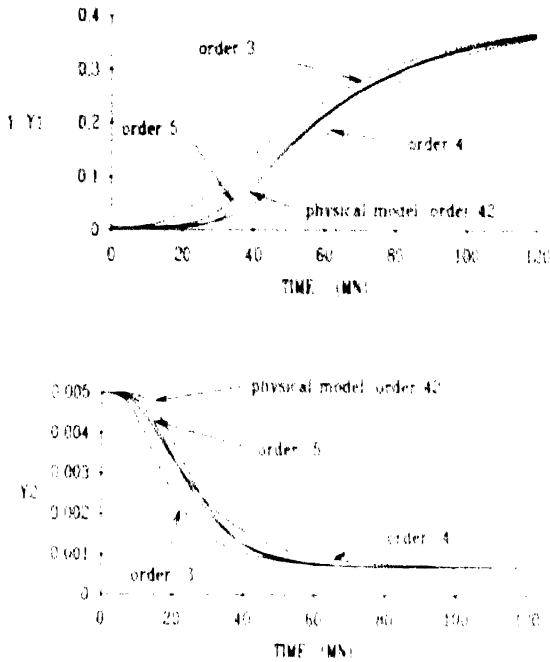


FIG. 4 Simulated open-loop responses to a step change in the feed composition corresponding to several aggregated models of a depropanizer.

sections: the rectifying section (trays 2 to j_r) with its aggregation tray r ; the feed section (trays $j_r + 1$ to $j_f - 1$) with its aggregation tray j_f ; the stripping section (trays j_f to $n - 1$) with

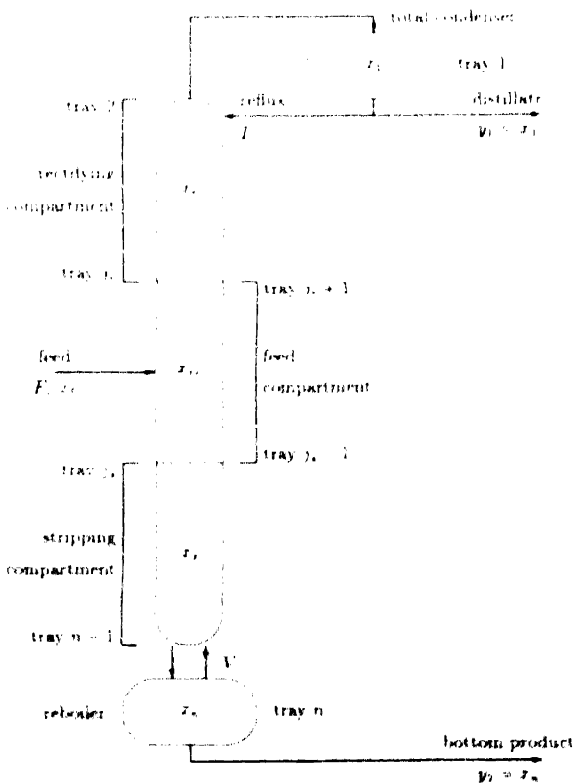


FIG. 5. The aggregation in 5 compartments.

its aggregation tray s ($2 < r < j_r < j_f < j_s < n$).

According to Theorem 3 (see also Remark 3), the reduced model is described by the differential-algebraic system

$$\begin{cases} \bar{H}_1 \dot{x}_1 = V k(x_2) - V x_1 \\ 0 = L x_{r-1} + V k(x_{r+1}) - L x_r - V k(x_r), \\ j = 2, \dots, r-1 \\ \bar{H}_r \dot{x}_r = L x_{r-1} + V k(x_{r+1}) - L x_r - V k(x_r) \\ 0 = L x_{j_r-1} + V k(x_{j_r+1}) - L x_{j_r} - V k(x_{j_r}), \\ j = r+1, \dots, j_f-1 \\ \bar{H}_{j_f} \dot{x}_{j_f} = L x_{j_f-1} + V k(x_{j_f+1}) - (L + F) x_{j_f} \\ - V k(x_{j_f}) - F z_f \\ 0 = (L + F) x_{j_f-1} + V k(x_{j_f+1}) - (L + F) x_{j_f} \\ - V k(x_{j_f}), \quad j = j_f+1, \dots, s-1 \\ \bar{H}_s \dot{x}_s = (L + F) x_{s-1} + V k(x_{s+1}) - (L + F) x_s \\ - V k(x_s) \\ 0 = (L + F) x_{s-1} + V k(x_{s+1}) - (L + F) x_s \\ - V k(x_s), \quad j = s+1, \dots, n-1 \\ \bar{H}_n \dot{x}_n = (L + F) x_{n-1} - (L + F - V) x_n - V k(x_n), \end{cases} \quad (17)$$

where

$$\bar{H}_1 = H_1, \quad \bar{H}_r = \sum_2^{j_r} H_j, \quad \bar{H}_{j_f} = \sum_{j_r+1}^{j_f} H_j, \\ \bar{H}_s = \sum_{j_f}^{n-1} H_j, \quad \bar{H}_n = H_n$$

The substitution of the algebraic equations into the 5 differential equations preserves the tri-diagonal structure of the original system (1) and gives the aggregated model

$$\begin{cases} \dot{x}_1 = f_1(x_1, x_r, L, V) \\ \dot{x}_r = f_r(x_1, x_r, x_{j_f}, L, V) \\ \dot{x}_{j_f} = f_{j_f}(x_r, x_{j_f}, x_s, L, V, F, z_f) \\ \dot{x}_s = f_s(x_{j_f}, x_s, x_n, L, V, F) \\ \dot{x}_n = f_n(x_s, x_n, L, V, F) \\ y_1 = x_1 \\ y_2 = x_n. \end{cases} \quad (18)$$

It results that $x^S = (x_1, x_r, x_{j_f}, x_s, x_n)$, and x^S corresponds to the liquid compositions of the remaining trays. (18) is called the control model and:

y_1 and y_2 are the outputs;

L and V are the control variables;

z_f is the perturbation;

F is a measured input;

The vector field f can be defined only with the

functions Y^m of Lemma 1,

$$\begin{aligned}
 \tilde{H}_1 f_1(x_1, x_r, L, V) &= -(V-L)x_1 + VY^{n-2}(L/V, x_1, x_r) - Lx_1 \\
 \tilde{H}_r f_r(x_1, x_r, x_{jr}, L, V) &= Lx_1 - VY^{n-2}(L/V, x_1, x_r) \\
 &\quad + VY^{n-1}(L/V, x_r, x_{jr}) - Lx_r \\
 \tilde{H}_{jr} f_{jr}(x_r, x_{jr}, x_i, L, V, F, z_f) &= Lx_r - VY^{n-1}(L/V, x_r, x_{jr}) \\
 &\quad + VY^{n-1}((L+F)/V, x_{jr}, x_i) \\
 &\quad - (L+F)x_{jr} \\
 \tilde{H}_i f_i(x_{jr}, x_i, x_n, L, V, F) &= (L+F)x_{jr} \\
 &\quad - VY^{n-1}((L+F)/V, x_{jr}, x_i) \\
 &\quad + VY^{n-1}((L+F)/V, x_i, x_n) \\
 &\quad - (L+F)x_i \\
 \tilde{H}_n f_n(x_i, x_n, L, V, F) &= (L+F)x_i \\
 &\quad - VY^{n-1}((L+F)/V, x_i, x_n) \\
 &\quad - (L+F-V)x_n.
 \end{aligned} \tag{19}$$

From now on, we shall only work with this control model. For obvious notational reasons, x and f , previously used for the physical model, remain unchanged since no ambiguity is possible: $x = x^0 = (x_1, x_r, x_{jr}, x_i, x_n)$ and $f = (f_1, f_r, f_{jr}, f_i, f_n)$.

The proposed reduction preserves the open-loop behavior of the physical model.

Theorem 4. Assume that Assumptions 1 and 2 hold. Then we have the following assertions:

(i) For each initial condition x^0 in the open subset $[0, 1]^5$, the maximal solution of (18) is defined on $[0, +\infty[$ and remains in $[0, 1]^5$.

(ii) For each L, V, F and z_f , there exists a unique steady state \bar{x} in $[0, 1]^5$ of system (18). Moreover, if $k(x) > x$ for all $x \in [0, 1]$ then \bar{x} satisfies

$$1 > \bar{x}_1 > \bar{x}_r > \bar{x}_{jr} > \bar{x}_i > \bar{x}_n > 0.$$

(iii) If the functions L, V, F and z_f are constant and if x^0 lies in $[0, 1]^5$, then (18) is Lyapunov-stable and its solution converges to the unique steady-state associated to L, V, F and z_f ; moreover, for every $x \in [0, 1]^5$, the Jacobian matrix $\partial f / \partial x$ has real, distinct and negative eigenvalues.

The proof is a straightforward adaptation of that of Theorem 1 using Lemma 1. \square

Remark 4. Our method is of the same spirit as Benallou *et al.* (1986) or as España and Landau (1978) since it produces a reduced model of the

column where dynamics of trays are replaced by dynamics of sections of trays.

However, if we assume piecewise constant equilibrium, our model is not a special case of Benallou *et al.* (1986) compartmental model. Let us recall the Benallou *et al.* compartmentation assumption: "The dynamic behavior of a section of stages, or a compartment, can be represented by that of a single stage having the same holdup as the total compartment holdup and the composition of the compartment sensitive stage". The dynamic behavior of our aggregation tray does not satisfy this assumption since, even if steady state coincide with the ones of the compartment sensitive stage, transients may differ. More precisely, if we apply the reduction technique of Benallou *et al.* (1986) to the 5 compartments case as displayed on Fig. 5, we obtain a reduced model whose structure does not remain tridiagonal whereas our aggregated model does. This results from the application of the equations (24) to (32) in Benallou *et al.* (1986). One can verify that the differential equation corresponding to the feed compartment depends, in Benallou *et al.* (1986), on x_1 and x_n , the product compositions, whereas, in our aggregated model (18) it does not. On the other hand, as in Benallou *et al.* (1986), the steady states of the outputs $y_1 = x_1$ and $y_2 = x_n$ are preserved whatever the inputs L, V, F and z_f are.

Contrary to España and Landau (1978), we do not use bilinear approximations of (1) for which the calculated molar fractions do not necessarily remain in $[0, 1]$. Moreover, no identification of compartmental parameters is needed here.

Note also that the qualitative dynamic behavior of our aggregated model is deduced from the stability analysis of Rosenbrock (1962) whereas, for the reduced models of Benallou *et al.* (1986) and España and Landau (1978), this analysis remains to be done.

Remark 5. We can enrich our model by including hydraulic effects due to the liquid flowing down from tray to tray. These dynamics that have been neglected in the physical model (1), are, for most industrial columns, much faster than the slowest part of the dynamics of the compositions and include time constants similar to those of the fast part of the composition dynamics. Moreover, the application of the Tikhonov theorem to a physical model including such hydraulic effects will produce the same aggregated model (18) since for the slow hydraulic model the liquid holdups remain constant. This is why we presented the analysis without such hydraulic effects. To fix

ideas, the time constants on the depropanizer which is considered below, are around 5 mn for hydraulics and more than 30 mn for the compositions.

3 THE NONLINEAR CONTROL LAW

This section is devoted to the application of nonlinear perturbation rejection techniques on the control model (18). In most of the theoretical results that have been developed for nonlinear systems, the controls and the perturbations appear linearly (see Hirschorn, 1981; Isidori *et al.*, 1981; Isidori, 1989). However, the extension of these results to our problem, where this dependence is nonlinear, does not present major difficulties.

Firstly, we establish, on the control model (18), a local constructive existence result of a feedback law rejecting feed composition disturbances with stability. Then, we show how such a control law can be synthesized as output feedback when temperature measurements are considered. Finally, the implementation on a refinery depropanizer is presented.

3.1. Nonlinear disturbance rejection with stability

Theorem 5. Assume that assumption 2 holds and additionally that for all $x \in]0, 1[$, $k(x) > x$. Consider the dynamic system (18) and a steady-state, \bar{x} , corresponding to \bar{L} , \bar{V} , \bar{F} and \bar{z}_f satisfying assumption 1. The associated steady-state values of the outputs are denoted \bar{y}_1 and \bar{y}_2 . Then locally around \bar{x} , the following assertions hold true:

(i) There exists a unique control law (L, V) , solution of the nonlinear algebraic system

$$\begin{aligned} f_1(x_1, x_r, L, V) &= \phi_1(y_1, v_1) \\ f_n(x_1, x_n, L, V, F) &= \phi_2(y_2, v_2) \end{aligned} \quad (20)$$

depending on x_1 , x_r , x_s , x_n , F and on (v_1, v_2) which are the new control variables, the closed-loop dynamics of the outputs are

$$\begin{aligned} \dot{y}_1 &= \phi_1(y_1, v_1) \\ \dot{y}_2 &= \phi_2(y_2, v_2), \end{aligned}$$

where ϕ_1 and ϕ_2 are arbitrary smooth functions such that $\phi_1(\bar{y}_1, 0) = 0$ and $\phi_2(\bar{y}_2, 0) = 0$, and $v = (v_1, v_2)$ being the new control vector.

(ii) If the closed-loop dynamics, ϕ_1 and ϕ_2 , are chosen asymptotically stable, then the overall closed-loop system is asymptotically stable.

Remark 6. The additional assumption, $k(x) > x$ for all $x \in]0, 1[$, physically means that the first component is the light one. For a large majority

of binary mixtures whether $\forall x \in]0, 1[$, $k(x) > x$ or $\forall x \in]0, 1[$, $1 - k(x) > 1 - x$ and one can choose the light component as first one. Consequently, the same result holds true if $k(x) < x$ for all $x \in]0, 1[$.

Proof of (i). The control law is necessarily the solution of the nonlinear algebraic system (Isidori, 1989)

$$\begin{cases} \dot{y}_1 = x_1 = f_1(x_1, x_r, L, V) = \phi_1(y_1, v_1) \\ \dot{y}_2 = x_n = f_n(x_1, x_n, L, V, F) = \phi_2(y_2, v_2) \end{cases} \quad (21)$$

Let us prove that this system is locally invertible at the steady-state by using the implicit functions theorem.

With (19), (21) becomes

$$\begin{cases} Y'' z \left(\frac{L}{V}, x_1, x_r \right) - x_1 - \frac{\phi_1(y_1, v_1)}{V} = 0 \\ -\frac{L+F}{V} x_1 - \left(\frac{L+F}{V} - 1 \right) x_n \\ Y''^{n-1} \left(\frac{L+F}{V}, x_1, x_n \right) - \frac{\phi_2(y_2, v_2)}{V} = 0. \end{cases} \quad (22)$$

Denote Π^1 and Π^n the left-hand side functions of (22). At the steady-state, $x = \bar{x}$, $\phi_1 = \phi_2 = 0$ and (\bar{L}, \bar{V}) verifies (22). Let us compute the Jacobian matrix of (22). $k(x) > x$ for all $x \in]0, 1[$ implies that (11) holds true for Y''^{-2} and Y''^{n-1} . Consequently, at the steady state,

$$\begin{aligned} \frac{\partial \Pi^1}{\partial L} &= \frac{\bar{V}}{L} \frac{\partial \Pi^1}{\partial V} = \frac{1}{\bar{V}} \frac{\partial Y''^{-2}}{\partial L/V} > 0 \\ \frac{\partial \Pi^n}{\partial L} &= \frac{\bar{V}}{L+F} \frac{\partial \Pi^n}{\partial V} \\ &= -\frac{1}{\bar{V}} \left(\bar{y}_2 - \bar{x}_1 + \frac{\partial Y''^{n-1}}{\partial (L+F)/V} \right) < 0. \end{aligned}$$

This implies that, at the steady state,

$$\begin{vmatrix} \frac{\partial \Pi^1}{\partial L} & \frac{\partial \Pi^1}{\partial V} \\ \frac{\partial \Pi^n}{\partial L} & \frac{\partial \Pi^n}{\partial V} \end{vmatrix} < \begin{vmatrix} \frac{\partial \Pi^1}{\partial V} & \frac{\partial \Pi^n}{\partial L} \end{vmatrix},$$

since by assumption 1 we have $0 < \frac{\bar{L}}{\bar{V}} < 1$ and $0 < \frac{\bar{V}}{L+F} < 1$. The Jacobian matrix of system (22) is thus regular. The implicit function theorem ensures the existence and uniqueness, in a neighbourhood of $(\bar{x}_1, \bar{x}_r, \bar{x}_s, \bar{x}_n, \bar{L}, \bar{V}, \bar{F})$ and of $v_1 = v_2 = 0$, of the solution (L, V) of system (20). This solution depends regularly on $(x_1, x_r, x_s, x_n, F, \phi_1(y_1, v_1), \phi_2(y_2, v_2))$. \square

Proof of (ii). Since ϕ_1 and ϕ_2 are chosen asymptotically stable, it suffices to prove that the zero dynamics [see Byrnes and Isidori (1988)] is stable. This is shown directly by application of

the Routh-Hurwitz criterion [see Gantmacher (1966) for example] on the linear tangent approximation at the steady state.

According to (i), the unique static feedback is locally given by

$$\begin{aligned} L &= \Xi_1(x_1, x_r, x_i, x_n, F, \phi_1(x_1, v_1), \phi_2(x_n, v_2)) \\ V &= \Xi_2(x_1, x_r, x_i, x_n, F, \phi_1(x_1, v_1), \phi_2(x_n, v_2)), \end{aligned}$$

where Ξ_1 and Ξ_2 are continuously differentiable functions. The closed-loop dynamics are thus

$$\begin{aligned} \dot{x}_1 &= \phi_1(x_1, v_1) \\ \dot{x}_r &= f_r(x_1, x_r, x_i, x_n, F, \phi_1(x_1, v_1), \phi_2(x_n, v_2)), \\ \Xi_1(x_1, x_r, x_i, x_n, F, \phi_1(x_1, v_1), \phi_2(x_n, v_2)), \\ \Xi_2(x_1, x_r, x_i, x_n, F, \phi_1(x_1, v_1), \phi_2(x_n, v_2))) \\ \dot{x}_i &= f_i(x_r, x_i, x_n, \Xi_1(x_1, x_r, x_i, x_n, F, \\ \phi_1(x_1, v_1), \phi_2(x_n, v_2))), \\ \Xi_2(x_1, x_r, x_i, x_n, F, \\ \phi_1(x_1, v_1), \phi_2(x_n, v_2)), F, z_f) \\ \dot{x}_n &= f_n(x_i, x_n, \Xi_1(x_1, x_r, x_i, x_n, F, \phi_1(x_1, v_1), \\ \phi_2(x_n, v_2)), \Xi_2(x_1, x_r, x_i, x_n, F, \phi_1(x_1, v_1), \\ \phi_2(x_n, v_2)), F) \\ \dot{x}_n &= \phi_2(x_n, v_2). \end{aligned} \quad (23)$$

The dynamics of x_1 and x_n are decoupled and stable by assumption. Consequently, the stability, around the steady-state \bar{x} , is ensured if the zero dynamics, obtained by setting ϕ_1 to 0, ϕ_2 to 0, x_1 to \bar{y}_1 and x_n to \bar{y}_2 in (23), is asymptotically stable. That is, if

$$\begin{aligned} \dot{x}_r &= f_r(\bar{y}_1, x_r, x_i, \bar{y}_2, \Xi_1(\bar{y}_1, x_r, x_i, \bar{y}_2, F, 0, 0), \\ \Xi_2(\bar{y}_1, x_r, x_i, \bar{y}_2, F, 0, 0)) \\ \dot{x}_i &= f_i(x_r, x_i, x_n, \Xi_1(\bar{y}_1, x_r, x_i, \bar{y}_2, F, 0, 0), \\ \Xi_2(\bar{y}_1, x_r, x_i, \bar{y}_2, F, 0, 0), F, z_f) \\ \dot{x}_n &= f_n(x_i, x_n, \bar{y}_2, \Xi_1(\bar{y}_1, x_r, x_i, \bar{y}_2, F, 0, 0), \\ \Xi_2(\bar{y}_1, x_r, x_i, \bar{y}_2, F, 0, 0), F) \end{aligned} \quad (24)$$

is asymptotically stable.

In the remaining part of this proof, all functions are evaluated at the steady-state. For simplicity's sake, we also denote $\frac{\partial L}{\partial x_i}$ in place of $\frac{\partial \Xi_1}{\partial x_i}$, and the same for V . The equations (22) defining the control law become

$$\begin{aligned} Y^{n-1} \left(\frac{L}{V}, \bar{y}_1, x_r \right) - \bar{y}_1 &= 0 \\ \frac{L+F}{V} x_r - \left(\frac{L+F}{V} - 1 \right) \bar{y}_2 \\ - Y^{n-1} \left(\frac{L+F}{V}, x_r, \bar{y}_2 \right) &= 0 \end{aligned}$$

when $\phi_1 = \phi_2 = 0$. The inequalities (11) imply that

$$\begin{aligned} \frac{V}{L+F} \frac{\partial L}{\partial x_r} = \frac{\partial V}{\partial x_r} < 0, \quad \frac{\partial L/V}{\partial x_r} < 0, \\ \frac{\partial(L+F)/V}{\partial x_r} = 0, \end{aligned} \quad (25)$$

$$\frac{\partial L/V}{\partial x_1} = 0, \quad \frac{\partial L}{\partial x_i} = \frac{L}{V} \frac{\partial V}{\partial x_i} > 0, \quad \frac{\partial(L+F)/V}{\partial x_i} < 0. \quad (26)$$

Using (19), the closed-loop system (24) can be rewritten as follows:

$$\begin{aligned} \bar{H}_r \dot{x}_r &= -g_1 + g_r, \\ \bar{H}_i \dot{x}_i &= F z_f - g_r - g_i, \\ \bar{H}_n \dot{x}_n &= -g_n + g_i, \end{aligned} \quad (27)$$

where g_1 , g_r , g_i and g_n are the following functions of (x_r, x_i, x_n)

$$\begin{aligned} g_1(x_r, x_i) &= (V(x_r, x_i) - L(x_r, x_i)) \bar{y}_1 \\ g_r(x_r, x_i, x_n) &= V(x_r, x_i) Y^{n-1} \left(\frac{L}{V}(x_r, x_i, x_n) \right) \\ &\quad - L(x_r, x_i) x_r \\ g_i(x_r, x_i, x_n) &= (L(x_r, x_i) + F) x_i \\ &\quad - V(x_r, x_i) Y^{n-1} \left(\frac{L+F}{V}(x_r, x_i, x_n) \right) \\ g_n(x_r, x_i) &= (L(x_r, x_i) + F - V(x_r, x_i)) \bar{y}_2. \end{aligned}$$

Let us now compute the tangent linearization to (27). (26) and $g_r - g_1 = 0$ at the steady-state

imply that $\frac{\partial}{\partial x_i} (g_r - g_1) = 0$. Symmetrically, we have $\frac{\partial}{\partial x_r} (g_i - g_n) = 0$. Consequently the matrix

A corresponding to the tangent model of the zero dynamics (27) has the form

$$A = \begin{pmatrix} \frac{a_1 + a_2}{\bar{H}_r} & \frac{a_1}{\bar{H}_r} & 0 \\ -\frac{a_2 + b_2}{\bar{H}_i} & -\frac{a_1 + b_1}{\bar{H}_i} & -\frac{a_4 + b_4}{\bar{H}_n} \\ 0 & \frac{b_1}{\bar{H}_r} & \frac{b_1 + b_4}{\bar{H}_n} \end{pmatrix},$$

where:

$$\begin{aligned} a_1 &= -\frac{\partial g_1}{\partial x_r}, & a_2 &= \frac{\partial g_r}{\partial x_r}, & a_3 &= \frac{\partial g_r}{\partial x_i}, & a_4 &= \frac{\partial g_r}{\partial x_n}, \\ b_1 &= -\frac{\partial g_n}{\partial x_i}, & b_2 &= \frac{\partial g_i}{\partial x_i}, & b_3 &= \frac{\partial g_i}{\partial x_r}, & b_4 &= \frac{\partial g_i}{\partial x_n}. \end{aligned}$$

The eigenvalues λ of A are solution of

$$\begin{vmatrix} \frac{a_1 + a_2}{\bar{H}_r} - \lambda & \frac{a_3}{\bar{H}_r} & 0 \\ -\frac{a_2 + b_2}{\bar{H}_{rr}} & -\frac{a_3 + b_3}{\bar{H}_{rr}} - \lambda & -\frac{a_4 + b_4}{\bar{H}_{rr}} \\ 0 & \frac{b_3}{\bar{H}_r} & \frac{b_1 + b_4}{\bar{H}_r} - \lambda \end{vmatrix} = 0,$$

or equivalently of

$$\begin{vmatrix} a_1 + a_2 - \bar{H}_r \lambda & a_3 & 0 \\ a_1 - b_2 - \bar{H}_r \lambda & -\bar{H}_{rr} \lambda & b_1 - a_4 - \bar{H}_r \lambda \\ 0 & b_3 & b_1 + b_4 - \bar{H}_r \lambda \end{vmatrix} = 0.$$

We have

$$\begin{aligned} a_1 + a_2 &= \frac{\partial(L - V)}{\partial x_r} \bar{y}_1 + \frac{\partial(VY^{m-r-1} - Lx_r)}{\partial x_r} \\ &= \frac{L + F - V}{L + F} \bar{y}_1 \frac{\partial L}{\partial x_r} + \frac{VY^{m-r-1} - (L + F)x_r}{L + F} \frac{\partial L}{\partial x_r} \\ &\quad + V \left(\frac{\partial Y^{m-r-1}}{\partial L/V} \right) \frac{\partial L/V}{\partial x_r} + V \frac{\partial Y^{m-r-1}}{\partial x_r} \cdot L \\ &= \frac{F(\bar{y}_1 - x_r)}{L + F} \frac{\partial L}{\partial x_r} + V \left(\frac{\partial Y^{m-r-1}}{\partial L/V} \right) \frac{\partial L/V}{\partial x_r} \\ &\quad + V \frac{\partial Y^{m-r-1}}{\partial x_r} \cdot L, \end{aligned}$$

(use $g_1 = g_r$ at the steady-state, (25), (10) with $m = j_f - r - 1$ and x_r playing the role of \bar{x}_m , $\bar{y}_1 > \bar{x}_r$). Consequently $a_1 + a_2 < 0$. Similar computations give

$$\begin{aligned} a_3 &> 0, \quad a_1 - b_2 < 0, \quad b_1 - a_4 < 0, \\ b_3 &> 0, \quad b_1 + b_4 < 0. \end{aligned}$$

Denote

$$\alpha_1 = -\frac{a_1 + a_2}{\bar{H}_r}, \quad \alpha_2 = -\frac{a_1 - b_2}{\bar{H}_r}, \quad \bar{H}_r$$

and

$$\beta_1 = -\frac{b_1 + b_4}{\bar{H}_r}, \quad \beta_2 = -\frac{b_1 - a_4}{\bar{H}_r}, \quad \beta_3 = \frac{b_3}{\bar{H}_{rr}}.$$

$\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2$ and β_3 are positive. The characteristic polynomial of A is then

$$\begin{aligned} \lambda^3 &+ (\alpha_1 + \beta_1 + \alpha_3 + \beta_3)\lambda^2 \\ &+ (\alpha_1\beta_1 + \alpha_3\alpha_2 \\ &+ \alpha_3\beta_1 + \beta_3\beta_2 + \beta_3\alpha_1)\lambda \\ &+ \alpha_2\alpha_3\beta_1 + \beta_2\beta_3\alpha_1. \end{aligned}$$

We have

$$\begin{aligned} \alpha_1 + \beta_1 + \alpha_3 + \beta_3 &> 0 \\ \alpha_1\beta_1 + \alpha_3\alpha_2 + \alpha_3\beta_1 + \beta_3\beta_2 + \beta_3\alpha_1 &> 0 \end{aligned}$$

$$\alpha_2\alpha_3\beta_1 + \beta_2\beta_3\alpha_1 > 0$$

$$\begin{aligned} (\alpha_1 + \beta_1 + \alpha_3 + \beta_3)(\alpha_1\beta_1 + \alpha_3\alpha_2 + \alpha_3\beta_1 + \beta_3\beta_2 \\ + \beta_3\alpha_1) > \alpha_2\alpha_3\beta_1 + \beta_2\beta_3\alpha_1. \end{aligned}$$

The desired stability result follows from the Routh-Hurwitz criterion. \square

Remark 7. The preceding result is only a local stability result. Simulations show that stronger stability properties should be expected, but we have no proof of them. On the other hand, it should be noticed that, in (i) of Theorem 5, ϕ_1 and ϕ_2 can be chosen in order to follow an arbitrary slow reference model. It results that set-point changes are easily handled by such a control technique. An integral action can be added through v_1 and v_2 in order to remove static offsets between the control model and the real column.

Remark 8. Equation (20) shows that x_{rr} is not required to compute the control law. This results from the special tridiagonal structure of the reduced model (18) and from the disturbance rejection method. More precisely, one can show, using the results of Isidori (1989), that the characteristic numbers of the outputs y_1 and y_2 with respect to the control are zero which means that the control variables affect the first time derivatives of y_1 and y_2 . The perturbation rejection method consists in making unobservable all the state variables which are affected by the perturbations and which need a larger number of derivations to affect y_1 and y_2 . Indeed, otherwise they would reintroduce the effects of the perturbations in the outputs. This explains why the feedback makes x_r, x_{rr} and x_r unobservable through y_1 and y_2 and why x_{rr} , which affects the outputs after 2 time derivations, does not appear in the control law. Note that the same results would hold true with more compartments and that the feedback scheme would remain the same. Namely, this would produce only an increase of the number of unobservable state variables by a feedback law still depending on 4 state variables.

Remark 9. The closed-loop analysis has been done under the condition that $k(x) > x$ for all $x \in [0, 1]$. Azeotropic mixtures [see Prausnitz *et al.* (1980) for example] do not satisfy such conditions. Nevertheless, the same result can in fact be proven by assuming that $k(x) > x, \forall x \in [0, a[$ and $k(x) < x, \forall x \in]a, 1[$ where $a \in [0, 1[$ is the azeotropic composition ($k(a) = a$). The reason is that the steady-state compositions are all on the same side of a if $z_f \neq a$.

3.2. Synthesis by output feedback

The control law of Theorem 5 depends on $(x_1, x_r, x_r, x_{rr}, F)$ and on the reference model

(21). In practice, the product compositions (x_1, x_n) and the feed flow F are measured. But the average compositions (x_r, x_s) in the rectifying and stripping compartments are not measured. Nevertheless, (x_r, x_s) can be estimated by means of well placed temperatures which leads to a reasonable approximation of the control law.

With the equilibrium function k , the equations of the thermodynamic equilibrium also give the temperature T on each tray as a function of its liquid composition x (see Prausnitz *et al.*, 1980): $T = \Theta(x)$. We now prove that T_r and T_s , the temperatures on the aggregation trays r and s , can be considered, at the order zero in ϵ , as functions of x_r and x_s respectively, and thus can be seen as additional outputs for system (18).

Let us return to the compartment of m trays of 2.3.1. Consider the temperature of the aggregated tray j_a , \bar{T}_{j_a} , and its liquid composition, \bar{x}_{j_a} . We have: $\bar{T}_{j_a} = \Theta(\bar{x}_{j_a})$. \bar{x}_{j_a} depends on the slow variable \bar{x}^S but also on the fast variables \bar{x}^F as follows:

$$\bar{x}_{j_a} = \frac{\bar{x}^S + \epsilon \sum_{i \neq j_a} \alpha_i \bar{x}_i^F}{1 + \epsilon \sum_{i \neq j_a} \alpha_i}$$

At the order 0 in ϵ , $\bar{x}_{j_a} = \bar{x}^S$. Consequently, $\bar{T}_{j_a} = \Theta(\bar{x}^S) + O(\epsilon)$. For the overall column, we have similarly

$$T_r = \Theta(x_r) + O(\epsilon) \quad \text{and} \quad T_s = \Theta(x_s) + O(\epsilon).$$

This implies that x_r and x_s can be estimated via temperatures on aggregation trays r and s . Moreover, these estimates do not contain fast components at the order 0 in ϵ . Thus the feedback law where x_r and x_s are replaced by $\Theta^{-1}(T_r)$ and $\Theta^{-1}(T_s)$ neither destabilizes the neglected fast dynamics nor the aggregated closed-loop system. Otherwise stated, the proposed synthesis by output feedback does not remix the time scales of the original system [for an extended discussion see Kokotovic (1984) or Marino and Kokotovic (1988)].

To summarize, the control law can be computed with the following online measurements:

- The product compositions, x_1 and x_n ;
- The rectifying temperature on tray r , T_r ;
- The stripping temperature on tray s , T_s ;
- The feed flow, F .

We have observed that the position of the aggregation trays r and s does not require great precision. For the depropanizer described below, several choices of r and s have been explored. The corresponding simulations show only slight differences. An important byproduct of this

approach is that, when the setpoints remain unchanged for a long period, the measurements of the product composition ($1 - y_1, y_2$) are not necessary. Consequently, failures on these measurements can appear without significantly affecting the behavior of the control law.

Notice that it results from Remark 8 that the number of on-line measurements (temperatures) does not depend on the number of compartments of the reduced model.

3.3. Implementation on a refinery depropanizer

The control law of Theorem 5 has been implemented on a refinery depropanizer (a binary column splitting a mixture of propane and butane into two products: the top product, essentially propane, and the bottom product, essentially butane). This depropanizer has the following characteristics:

- 42 theoretical trays ($n = 42$), feed on tray 21 ($j_f = 21$).
- The holdup profile is as follows: on tray 1 (reflux drum): 60 kmol; on trays 2 to 41: 2 kmol; on tray 42 (bottom): 30 kmol.
- The top pressure: 15 bar.
- A typical steady-state is: saturated liquid feed flow of 5 kmol/min with a propane molar fraction around 0.35, a reflux flow of 5 kmol/min, a reboiler vapor outflow of 7 kmol/min, product purities of 0.5% butane in the top product, and of 0.5% propane in the bottom product.

For this column, the modeling assumptions are valid, and the hypothesis of Theorems 1, 3, 4 and 5 are satisfied. The real-time control law depends on:

- The molar fraction of butane in the top product;
- The molar fraction of propane in the bottom product;
- The rectifying temperature on tray 11;
- The stripping temperature on tray 33;
- The feed flowrate;
- The top pressure as a parameter in the thermodynamic calculations.

It should be mentioned that the two composition measurements are obtained with a delay greater than 5 min. The first control L , the reflux flow, is directly measured and regulated. The second control V , the reboiler vapor outflow, is proportional to the reboiler duty which is measured and regulated. The thermodynamic model (the functions k and Θ), used to represent the binary mixture propane-butane, is borrowed from Soave (1972).

We now present records of real data relative

to this depropanizer. On Fig. 6, the variations over 10 hours of the product compositions ($1 - y_1$ and y_2), of the control variables (L and V) and of the measured input (F), are displayed. At time 0, the control law is switched on; the objectives are set to 0.5% butane in the distillate (setpoint of $1 - y_1$) and to 0.3% propane in the bottom product (setpoint of y_2). These data suggest two comments. Firstly, though important initial offsets exist between outputs and setpoints, the objectives are reached in 5 hours (the time-constants of the linear first-order reference models are around 2 hours). Secondly, the outputs are only slightly modified in spite of severe variations of the feed flow F (more than 40% in 15 min). This demonstrates that the nonlinear control law works in a large range of operating conditions and rejects the perturbations asymptotically.

To conclude this section, we have compared our control technique with the following classical SISO method:

- The reflux flow L is proportional to the feed flow F with a gain slowly adapted by PI controller depending on the top composition y_1 ;
- The reboiler duty proportional to V is controlled through PI action depending on the stripping temperature T_s .

Figure 7 displays on-line data relative to the same depropanizer where the nonlinear control is removed after $t = 360$ min and is replaced, in the absence of feed perturbations, by the SISO control described above. The bottom quality remains acceptable whereas the top quality

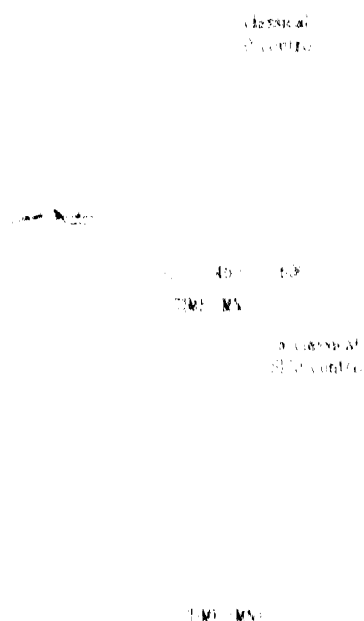


Fig. 7. Plant data, comparison between the aggregated nonlinear control and a classical SISO control technique on a refinery depropanizer

slowly becomes out of specifications. This SISO controls the quality of one of the two products in a quite satisfactory way, but it is unable to control simultaneously the top and bottom product qualities.

4. DISCUSSION

All the dynamic simulations presented in this section correspond to the refinery depropanizer of the preceding section and are obtained via the dynamic simulator SPEEDUP [User Manual (1988)]. Numerical integrations start from the same steady-state characterized by:

- Feed flowrate 5 kmol/min, reflux flowrate 5.088 kmol/min, vapor leaving the reboiler 6.957 kmol/min;
- Column pressure 15 bar;
- Feed composition: 2.5% ethane, 35% propane, 60% *n*-butane, 2.5% *n*-pentane.

During the first 10 min, the feed compositions change to the new values: 2.5% ethane (unchanged), 20% propane, 75% *n*-butane, 2.5% *n*-pentane (unchanged). After that, all the entries remain unchanged. The thermodynamic model used to represent the liquid-vapor equilibria is the Soave model (1972). The open-loop responses can be seen on Fig. 4 where a similar perturbation of the feed composition is introduced.

The set-points are: 0.5% *n*-butane in the top product, 0.5% propane in the bottom product. Notice that apart from the two key components (propane, *n*-butane), we add two other secondary ones, present in practice in small quantities.

In the control law of Theorem 5, the output

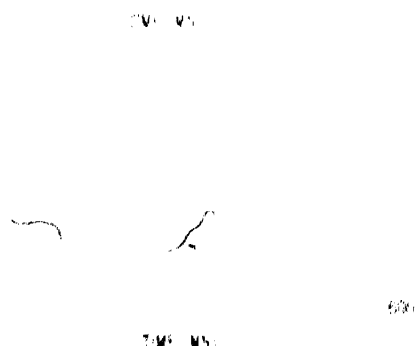


Fig. 6. Plant data, implementation of the aggregated nonlinear control on a refinery depropanizer

dynamics (the functions ϕ_1 and ϕ_2) are arbitrary stable dynamics. For the simulation tests, we choose

$$\phi_1(y_1, v_1) = \alpha \quad \phi_2(y_2, v_2) = \frac{v_2 - y_2}{\alpha}$$

where $\alpha = 10$ min is constant, $v_1 = 0.995$ is the top set-point and $v_2 = 0.005$ the bottom one (the component chosen to write the balance equations is the *n*-butane).

Our control technique is now compared with other linear and nonlinear nonaggregated methods. We do not consider errors on the state measurements since an observer ought to be designed in each case and the comparisons would be hazardous. Only robustness versus delays is studied.

4.1. Why nonlinear control?

We compare, by simulation, the performance of the nonlinear control law of Theorem 5 with the linear geometric control law of Takamatsu *et al.* (1979). This linear control law is the solution of the linear system

$$\begin{aligned} a_{1,2}b_{2,1}\delta L + a_{1,2}b_{2,2}\delta V &= \delta x_1 - (a_{1,1}^2 + a_{1,2}a_{2,1}) \\ &\quad \times \delta x_1 - a_{1,2}(a_{1,1} + a_{2,2}) \\ &\quad \times \delta x_2 - a_{1,2}a_{2,1}\delta x_3 \\ b_{n,1}\delta L + b_{n,2}\delta V &= \delta x_n - a_{n,n-1}\delta x_{n-1} \\ &\quad - a_{n,n}\delta x_n - b_{n,1}\delta F, \end{aligned}$$

where

$(a_{i,j})_{1 \leq i,j \leq n}$ and $(b_{i,j})_{1 \leq i,j \leq n-1,2}$ are obtained by linearization at the steady-state;

$(\delta x_i)_{1 \leq i \leq n}$ are the deviations of the state x (see system (1)), $(\delta L, \delta V)$ the deviations of the control (L, V) and δF the deviation of the measured input F ;

δx_1 and δx_n are the closed-loop output dynamics, chosen linear and stable:

$$\begin{aligned} \delta x_1 &= -\left(\frac{1}{\theta} + \frac{1}{2\theta}\right)\delta x_1 - \frac{1}{2\theta^2}\delta x_2 \\ \delta x_n &= -\frac{\delta x_n}{\theta}, \end{aligned} \quad (28)$$

with $\theta = 5$ min, a constant and $\delta x_1 = a_{1,1}\delta x_1 + a_{1,2}\delta x_2$.

In simulations, we suppose that the part of the state required for control is measured directly (the composition of propane on tray 1 and 2, used for the control, are calculated by 1 minus the true compositions of *n*-butane on these trays).

Figure 8 corresponds to the closed-loop output responses to feed composition perturbations, firstly when the state is perfectly known and secondly when a measurement delay of 5 min is

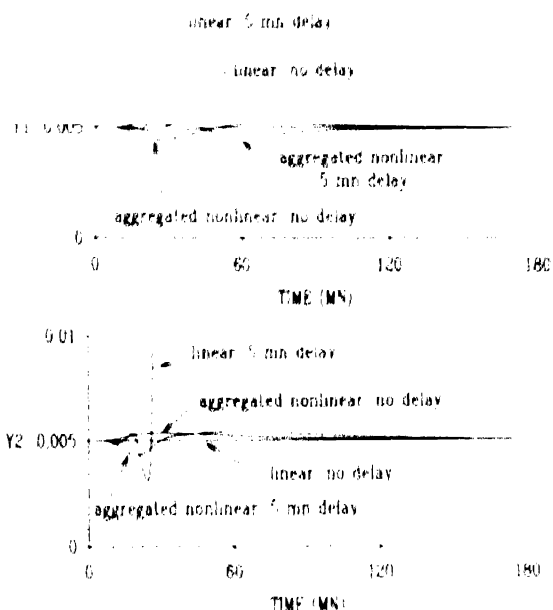


FIG. 8. Simulation data, comparison between the aggregated nonlinear control law and the linear geometric control law (depropanizer, step change of the feed composition)

introduced. This delay corresponds to the resident time in the chromatograph, the composition sensor generally used for a depropanizer. The linear control law works better than the nonlinear aggregated one if the measurements are perfect and without delays. But, the nonlinear aggregated control law is less sensitive to measurement delays, whereas the linear one blows up in their presence.

In other simulations, we have observed that the parameter θ of (28) must be carefully chosen: if θ is too large, for instance greater than 10 min, the linear control law destabilizes the column. The gains of the control law must be large enough to maintain the linear model in its validity region that seems to be small.

4.2. Why aggregation?

We have also compared the nonlinear aggregated control law with the nonlinear control law rejecting feed composition disturbances in the system (1).

We have proven in Levine and Rouchon (1986) that, at the steady-state, the nonlinear control law of Gauthier *et al.* (1983) is singular. In order to bypass this difficulty, one can look for nonmaximal invariant distributions (see Isidori, 1989). Since on the singularity $x_1 = k(x_2)$, the reader can verify that this leads to rejecting the perturbation on the new output functions $(y_1 = k(x_2), y_2 = x_n)$. That is, we change the top output (x_1) to the propane composition of the vapor leaving tray 2 ($k(x_2)$). Clearly, since $H_1\dot{x}_1 = V(k(x_2) - x_1)$, if x_2 is constant then x_1 is also constant. With these new outputs the control law is regular near the

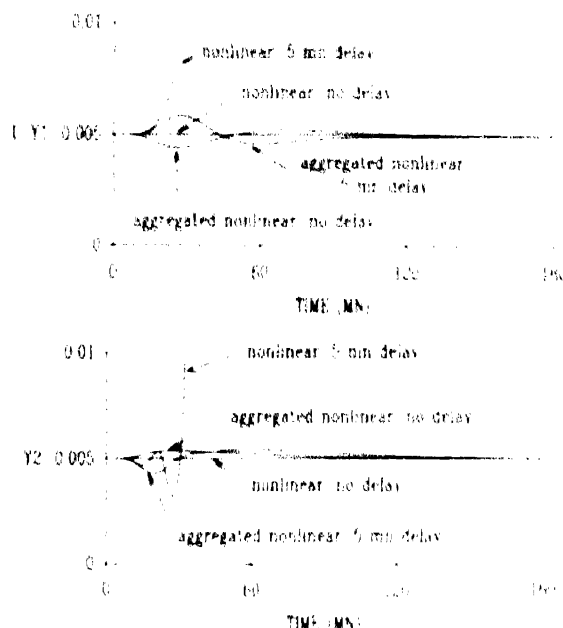


FIG. 9. Simulation data: comparison between the aggregated nonlinear control law and the complete nonlinear one elaborated on the physical model (depropanizer, step change of the feed composition)

steady-state and given by

$$\left\{ \begin{aligned} & \frac{dk}{dx}(x_2) \frac{x_1 - x_2}{H_2} L + \frac{dk}{dx}(x_2) \frac{k(x_1) - k(x_2)}{H_2} V \\ &= \frac{v_1 - k(x_2)}{\beta} \\ & \frac{x_{n-1} - x_n}{H_n} L + \frac{x_n - k(x_n)}{H_n} V \\ &= \frac{v_2 - x_n}{\beta} + \frac{x_n - x_{n-1}}{H_n} F, \end{aligned} \right. \quad (29)$$

where we chose linear stable output dynamics with $\beta = 10$ min ($v_1 = 0.995$ and $v_2 = 0.005$).

As for the linear geometric control, two simulations have been made: the first one with perfect state measurements and no delay, the second one with a measurement delay of 5 min (see Fig. 9). The complete nonlinear control law works perfectly without delays. But the linear geometric control destabilizes the column in the presence of delays. Instead of the linear geometric control, the choice of the parameter β in (29) is not important: with perfect state knowledge, stability is ensured if $\beta > 0$.

5. CONCLUSIONS

We have applied the nonlinear perturbation rejection techniques to an aggregated model of distillation columns. The obtained control law is shown to be robust and simple to implement. It is actually working on two depropanizers and two debutanizers of ELF-FRANCE.

Another by-product of our control technique concerns the design of the instrumentation of the columns. The proposed control law uses

intermediate temperatures; their positions can be adjusted in order to give the best closed-loop responses (composition and noise sensitivities).

The same technique can be extended to other counter-current separation processes, as in Duchêne (1988), where counter-current mixer-settler extractors are studied and similar results are obtained. Moreover, the extensions of these results to more complex distillation columns (more than two components, several feeds, side products, networks of columns) seems to be reasonable.

The implementation on the refinery depropanizer has shown that controlling product qualities has several industrial interests, in particular:

- Energy savings: the purity margins that the operator maintains with the final product specification can be reduced, as well as the associated energy consumption (for the depropanizer, savings are greater than 15%);
- Productivity gains: the internal fluid circulation is also reduced, the trays are less flooded and the feed flowrate can be increased (for the depropanizer, productivity gains are greater than 15%);
- Process flexibility.

Acknowledgements.—The authors are indebted to Professor H. Renon for numerous fruitful discussions and to Y. Creff for useful suggestions concerning the proof of Lemma 2 and the spectral property of Theorems 1 and 4. The authors wish also to thank ELF-FRANCE of the group ELF-AQUITAINE for supporting this research program, and especially F. Djenah of the Centre de Recherche d'ELF à Solaize, A. Douaud and J. P. Beauchêne of the ELF Refinery of Donges for collaborating in the implementation on the depropanizer. The financial support of the E.E.C. (Non Nuclear Energy R. & D. Programme) is also gratefully acknowledged.

REFERENCES

- Agarwal, M. and D. E. Seborg (1987). A multivariable nonlinear self-tuning controller. *AIChE J.* **33**, 1376–1386.
- Alsop, A. W. and T. F. Edgar (1987). Nonlinear control of a high purity distillation column by the use of partially linearized control variables. *AIChE Spring National Meeting*, Houston, Texas.
- Arnold, V. (1974). *Equations Différentielles Ordinaires*. Mir, Moscow.
- Benallou, A., D. E. Seborg and D. A. Mellichamp (1986). Dynamic compartmental models for separation processes. *AIChE J.* **32**, 1067–1078.
- Byrnes, C. I. and A. Isidori (1988). Local stabilization of minimum-phase nonlinear systems. *Syst. Control Lett.* **11**, 9–17.
- Duchêne, P. (1988). Simulation dynamique et contrôle d'une cascade de mélangeurs-décanteurs. Rapport de DEA, École des Mines de Paris, France, October 1988.
- España, M. and I. D. Landau (1978). Reduced order bilinear models for distillation columns. *Automatica*, **14**, 345–355.
- Fuente, C. and W. L. Luyben (1983). Control of high-purity distillation columns. *Ind. Engng. Chem. Proc. Des. Dev.*, **22**, 361–366.
- Gallun, S. E. and C. D. Holland (1982). Gear's procedure for the simultaneous solution of differential and algebraic equations with applications to unsteady state distillation problems. *Computers Chem. Engng.* **6**, 231–244.

- Gantmacher, F. R. (1966). *Théorie des Matrices: Tome 2*. Dunod, Paris.
- Gauthier, J. P., G. Bornard, S. Bacha and M. Idir (1983). Rejet de Perturbations pour un Modèle Non Linéaire de Colonne à Distiller. *Developpement et Utilisation d'Outils et Modèles Mathématiques en Automatique, Analyse des Systèmes et Traitement du Signal*, CNRS, Paris.
- Georgakis, C. (1986). On the use of extensive variables in process dynamics and control. *Chem. Engng Sci.*, **41**, 1471-1484.
- Georgiou, A., C. Georgakis and W. L. Luyben (1988). Nonlinear dynamic matrix control for high-purity distillation columns. *AIChE J.*, **34**, 1287-1298.
- Hirschorn, K. M. (1981). (A, B)-invariant distributions and disturbance decoupling problem of nonlinear systems. *SIAM J. Control Optimiz.*, **19**, 1-19.
- Hunt, L. R., R. Su and G. Meyer (1983). *Design for Multi-input Nonlinear Systems* (Volume 268 of R. W. Brockett, R. S. Millman and H. J. Sussman (Eds) *Differential Geometric Control Theory*).
- Isidori, A. (1989). *Nonlinear Control Systems*. Communications and Control Engineering Series. Springer, Berlin.
- Isidori, A., A. Krener, C. Gori-Giorgi and S. Monaco (1981). Nonlinear decoupling via feedback. *IEEE Trans. Auto Control*, **26**, 331-345.
- Jakubczyk, B. and W. Respondek (1980). On linearization of control systems. *Bull. Acad. Pol. Sci. Ser. Sci. Math.*, **28**, 517-522.
- Kokotovic, P. V. (1984). Application of singular perturbation techniques to control problems. *SIAM Rev.*, **26**, 501-550.
- Krener, A. J. (1984). Approximate linearization by state feedback and coordinate change. *Syst. Control Lett.*, **5**, 181-185.
- Kummel, M. and H. W. Andersen (1987). Controller adjustment for improved nominal performance and robustness-II: Robust geometric control of distillation column. *Chem. Engng Sci.*, **42**, 2011-2023.
- Kwaalen, E., I. Neel and D. Tondeur (1985). Directions of quasi-static and energy transfer between phase in multicomponent open systems. *Chem. Engng Sci.*, **40**, 1191-1204.
- Lévine, J. and P. Rouchon (1986). Disturbances rejection and integral control of aggregated nonlinear distillation models. In *Proc. of the 7th Conf. on Analysis and Optimisation of Systems*. Lecture Notes in Control and Information Sciences. Springer, Berlin.
- Malesinski, W. (1965). *Azeotropy and Other Theoretical Problems of Vapour-Liquid Equilibrium*. Wiley-Interscience, New York.
- Marino, R. and P. V. Kokotovic (1988). A geometric approach to nonlinear singularly perturbed control systems. *Automatica*, **24**, 31-41.
- Morari, M. (1988). Advances in process control theory. *Chem. Engng Prog.*, **84**, 60-67.
- Prausnitz, J., T. Anderson, E. Grens, C. Eckert, R. Hsieh and J. O'Connell (1980). *Computer Calculations for Multi-component Vapor-Liquid and Liquid-Liquid Equilibria*. International Series in the Physical and Chemical Engineering Sciences, Prentice-Hall, Englewood Cliffs, NJ.
- Rosenbrock, H. H. (1962). A Liapunov function with applications to some nonlinear physical systems. *Automatica*, **1**, 31-53.
- Soave, G. (1972). Equilibrium constants from a modified Redlich-Kwong equation of state. *Chem. Engng Sci.*, **27**, 1197-1203.
- Speedup User Manual: Release 5.0. (1988). Prosys Technology Cambridge, U.K.
- Takamatsu, T., I. Hashimoto and Y. Nakai (1979). A geometric approach to multivariable system design of a distillation column. *Automatica*, **15**, 387-402.
- Tikhonov, A., A. Vasil'eva and A. Sveshnikov (1980). *Differential Equations*. Springer, Berlin, 1980.
- Van Winkle, M. (1967). *Distillation*. McGraw-Hill, 1967.
- Wonham, W. M. (1974). *Linear Multivariable Control: a Geometric Approach*. Springer, Berlin.

APPENDIX A. TWO RESULTS

Rosenbrock (1962) has proven the result given below (Theorem 6 of the appendix).

Rosenbrock's theorem Consider the differential system of dimension $p > 0$, $\dot{\xi} = \phi(\xi)$. Assume that $\xi = (\xi_i)_{i=1, \dots, p}$ belongs to Ω , a bounded, closed and convex subset of \mathbb{R}^p , and that $\phi = (\phi_i)_{i=1, \dots, p}$ with its partial derivatives are continuous functions of ξ . Suppose that:

- For each initial condition in Ω , the solution remains in Ω .
- For all $i \in \{1, \dots, p\}$, the function of ξ

$$\psi_i(\xi) = - \sum_{k=1}^p \frac{\partial \phi_k}{\partial \xi_i}(\xi) \quad (30)$$

is non-negative,

- For all i and k in $\{1, \dots, p\}$ such that $i \neq k$, $\partial \phi_k / \partial \xi_i \neq 0$.

(iv) Given any $i \in \{1, \dots, p\}$ for which $\psi_i = 0$, there exists $j \neq i$ in $\{1, \dots, p\}$ such that $\partial \phi_j / \partial \xi_i \neq 0$; if $\psi_i = 0$, there exists $k \in \{1, \dots, p\}$ different of i and j such that $\partial \phi_k / \partial \xi_i \neq 0$; if $\psi_k = 0$, then $i = j$; moreover, this process always leads in the end to some $l \in \{1, \dots, p\}$ for which $\psi_l \neq 0$.

Then there exists a unique steady-state in Ω , every solution starting in Ω converges to this steady-state and the function

$$\sum_{i=1}^p |\phi_i(\xi)|$$

is a Lyapunov function of the system.

We will also use the following lemma

Lemma 2 Consider $a = (a_i)$ and $b = (b_i)$, two real vectors of dimension $p > 0$, and the real $p \times p$ matrix $J = (J_{ij})$ constructed by means of a and b as follows:

- For $i = 2, \dots, p$, $J_{i, i-1} = a_{i-1}$.
- For $i = 1, \dots, p$, $J_{i, i} = -a_i - b_i$.
- For $i = 1, \dots, p-1$, $J_{i, i+1} = b_{i+1}$.
- For $i, j = 1, \dots, p$ such that $|i - j| > 1$, $J_{ij} = 0$.

If for all $i \in \{0, \dots, p\}$, $a_i > 0$ and $b_i > 0$, then the eigenvalues of J are distinct, real and negative.

The proof of this lemma is a straightforward application of a classical result relative to Jacobi's matrix [Gantmacher (1966, p. 99)].

APPENDIX B. THE TIKHONOV THEOREM

Consider the singularly perturbed system

$$\begin{aligned} \dot{x}^s &= f^s(x^s, x^f, u(t), w(t), \epsilon) - x^s(0) = x^{s0} \\ \epsilon \dot{x}^f &= f^f(x^s, x^f, u(t), w(t), \epsilon) - x^f(0) = x^{f0}, \end{aligned}$$

which admits continuous solution $(x^s(t, \epsilon), x^f(t, \epsilon))$ in $[0, T]$, $T > 0$ (f^s and f^f are continuously differentiable functions). The associated slow subsystem is

$$\begin{aligned} \dot{x}^s &= f^s(x^s, x^f, u(t), w(t), 0) - x^s(0) = x^{s0} \\ 0 &= f^f(x^s, x^f, u(t), w(t), 0) \end{aligned}$$

We suppose that it admits a continuous solution $(x_0^s(t), x_0^f(t))$ in $[0, T]$. For $t \in [0, T]$, the associated fast subsystem is

$$\frac{dx^f}{d\tau} = f^f(x_0^s(t), x^f(\tau), u(t), w(t), 0),$$

where $\tau = t/\epsilon$.

Tikhonov's theorem If, for each $t \in [0, T]$, the tangent linearization of the fast subsystem around $x_0^f(t)$ produces a stable linear system, and if x^{f0} belongs to the region of attraction of $x_0^f(0)$ then

$$\begin{aligned} \lim_{\epsilon \rightarrow 0^+} x^s(t, \epsilon) &= x_0^s(t) \\ \lim_{\epsilon \rightarrow 0^+} x^f(t, \epsilon) &= x_0^f(t) \end{aligned}$$

uniformly on all closed subsets of $]0, T[$. More details can be found for example in Tikhonov *et al.* (1980).

Drag-law Effects in the Goddard Problem*

P. TSOTRAS† and H. J. KELLEY‡

For a drag-law witnessing a sharp increase in the transonic region, a more complex switching structure than the classical full-singular-coasting sequence may occur during the optimal burning program for the vertical climb of a rocket.

Key Words—Aerospace trajectories, bang-bang control, optimal control, optimization, singular arcs, singular control

Abstract—Presently studied is the problem of maximizing the altitude of a rocket in vertical flight in a resisting medium, when the amount of propellant is specified, known as the Goddard problem. The case is studied in which the drag coefficient is a function of the Mach number, witnessing a sharp increase in the transonic region. Analysis shows the possibility of a more complex switching structure than the classical full-singular-coast sequence, with the appearance of a second full-thrust subarc in the transition from the subsonic to the supersonic region. Necessary conditions such as the Legendre-Clebsch condition for singular subarcs and the McDaniel-Powers condition for joining singular and non-singular subarcs were checked, and were found to be satisfied. It is shown that the results obtained depend heavily on the assumed form of the drag law, and on the magnitude of the upper bound on the thrust.

1. INTRODUCTION

THE PROBLEM of optimum thrust programming for maximizing the altitude of a rocket in vertical flight, for a given amount of propellant, has been extensively analyzed over the past sixty years. Briefly, we can refer to the pioneering work of Goddard (1919), Hamel (1927), Tsien and Evans (1951) and Leitmann (1956, 1957). However, as Leitmann (1963) first pointed out, the problem's solution continued to be far from complete, mainly due to difficulty arising from the

requirement that the mass be monotonically non-increasing. In this work, the possibility of a more complex sequence of subarcs for the case of sharp transonic drag-rise was suggested.

Solutions that meet this requirement have been obtained only in a few special cases, typified by the work of Miele and Cavoti (1958), and Miele (1955), who treated the cases of flight in vacuum and flight with a power law for drag, and later by Bryson and Ross (1958). Miele (1962), using a totally different approach, also proved the *sufficiency* of the optimal solution established by his predecessors, i.e. that the optimal burning program involves a rapid boost at the beginning of the flight, usually followed by a period of continuous burning (sustain phase) and ending with a zero-thrust period. Miele and Cicala (1956) were also the first to suggest the possibility of a more complex sequence of subarcs for the case of a general drag model.

One of the most complete works on Goddard's problem is perhaps the extensive treatment by Garfinkel (1963), who proved that with impulsive boosts in the velocity admitted, and for the case of a general drag model, the solution contains a finite number of such boosts in the transonic velocity region, and contains no coasting arcs except the terminal one.

As already has been established by previous researchers, the drag plays a significant role in the switching structure of the problem. In particular, it has been shown by Tsien and Evans (1951) and later by Miele (1955), that for the special case when drag is ideally zero, the variable-thrust subarc disappears from the extremal solution, which consequently reduces to subarcs flown with maximum engine output and coasting subarcs. Moreover, the approximation $C_D \approx \text{const.}$ may be of use at low altitudes, when the speed of optimum climb still belongs to

* Received 7 March 1989, revised 5 February 1990, received in final form 22 August 1990. The original version of this paper was presented at the 7th IFAC Workshop on Control Applications of Nonlinear Programming and Optimization which was held in Tbilisi, U.S.S.R. during June 1988. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor E. Kreindler under the direction of Editor H. Kwakernaak.

† Formerly with the Department of Aerospace and Ocean Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, U.S.A., currently with the School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN 47906, U.S.A.

‡ Department of Aerospace and Ocean Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, U.S.A. Dr H. J. Kelley passed away 9 February 1988.

the region of quasi-incompressible flow. As the altitude increases, both the velocity and the Mach number increase with such a rapidity that the hypothesis $\partial C_D / \partial M = 0$ is soon no longer satisfied, and a more accurate drag model should be used. Constant drag coefficient C_D is then replaced by a Mach-dependent drag coefficient featuring a sharp increase in the transonic region. When such a model is used, two optimal solutions for the singular surface arise: one in the subsonic-transonic region, and the other in the supersonic region of the velocity. For the case of level flight of a rocket-powered aircraft Miele and Cicala (1956) showed that another full-thrust subarc may occur during transition from the subsonic to supersonic region.

In the current work, Goddard's problem is examined with relaxed restrictions on the assumed drag characteristics of the rocket. The relaxed assumptions allow for switching structures that were previously not considered for the case of vertical climb. Invaluable insight to the problem was obtained via a transformation to a state-space of reduced dimension, where the problem becomes more tractable. The methodology can be applied to other singular optimal control problems too, in order to determine the possible optimal solution structure, i.e. the number and the relative position of singular and bang-bang subarcs.

2. PROBLEM FORMULATION

The vertical flight of a rocket obeys the following system of differential equations:

$$\begin{aligned}\dot{h} &= \dot{v} \\ \dot{v} &= \frac{\bar{T} - \bar{D}}{\bar{m}} - \bar{g} \\ \dot{\bar{m}} &= -\frac{\bar{T}}{\bar{c}}\end{aligned}\quad (1)$$

where \bar{h} , \bar{v} , \bar{m} are the altitude, velocity and mass respectively, \bar{T} denotes the engine thrust, \bar{D} denotes the aerodynamic drag, \bar{c} represents the exhaust velocity of the gases from the rocket engine, and \bar{g} is the gravitational acceleration. The second of the above equations simply states the force equilibrium along the flight path, the first equation is the kinematic relation between the altitude and the velocity, and the last states that the fuel consumption is proportional to the thrust. If we assume spherical earth with an inverse-square gravitational field, the above system of equations can be suitably nondimen-

sionalized using the following quantities:

$$\begin{aligned}h &= R_e \\ \dot{h} &= G^{-1/2} \dot{h}^{3/2} \\ \dot{v} &= G^{1/2} \dot{h}^{-1/2} \\ \bar{g} &= G \bar{h}^{-2} \\ \bar{m} &= m_0.\end{aligned}\quad (2)$$

Here R_e denotes the radius of the earth, G the gravitational constant, \bar{g} the acceleration due to gravity at the surface of the earth, and \bar{m} the launching mass of the vehicle. Using the above nondimensionalization factors, the equations of motion in nondimensionalized form become:

$$\begin{aligned}\dot{h} &= v \\ \frac{T - D}{m} &= h^{-2} \\ \dot{m} &= -\frac{T}{c}\end{aligned}\quad (3)$$

If S is the characteristic cross-section area of the vehicle, and ρ denotes the atmospheric density, then the aerodynamic drag is given by the quadratic formula:

$$D = C_D(M) S \frac{\rho v^2}{2} \quad (4a)$$

If in addition, we assume that the density of the atmosphere reduces exponentially with the altitude, the nondimensionalized form of the drag force becomes

$$D = C_D(M) b v^2 \exp(\beta(1 - h)) \quad (4b)$$

where the factor $b v^2 \exp(\beta(1 - h))$ is numerically equal to the product of the velocity head and the characteristic cross-section area of the aircraft, b , β are constants, and M is the Mach number defined as the ratio of the vehicle speed over the speed of sound. For simplicity it will be assumed that the speed of sound remains constant with altitude, an assumption which is actually valid only for stratospheric solutions. In (4) $C_D(M)$ is the zero-lift drag coefficient assumed to depend on the Mach number according to the following relationship:

$$C_D(M) = A_1 \tan^{-1}(A_2(M - A_3)) + A_4. \quad (5)$$

This formula generates a quick transition from one value of C_D in the subsonic region to another higher value of C_D in the supersonic region. The A_1 , A_2 , A_3 , A_4 are constants controlling when, and how fast, this transition takes place (Fig. 1).

The initial conditions are specified for the three state variables as h_0 , v_0 and m_0 . The final

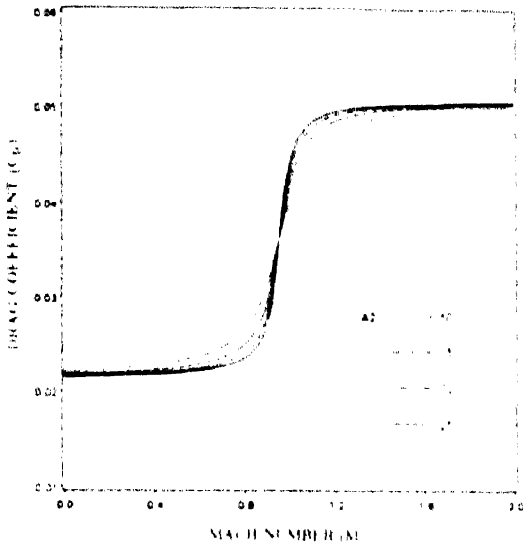


FIG. 1 Variation of the drag coefficient C_D with Mach number M .

value of the mass is also given as m_1 . The problem is to determine the optimum trajectory of a rocket in vertical flight, from an assigned initial position on the surface of the earth to the final position where the altitude reaches its maximum value, i.e. we want to maximize the altitude at the terminal time. Hence, the performance index is given by

$$J = h(t_f) \quad (6)$$

subject to the prescribed boundary condition

$$v(t_f) = 0 \quad (7)$$

and the dynamic equality constraints given by (3). The thrust is the control variable which is bounded according to the inequality:

$$0 \leq T \leq T_{\max} \quad (8)$$

The aerodynamic data and the vehicle's parameters, with the exception of the value for T_{\max} , were taken from the work of Zlatskiy and Kiforenko (1983), and their nondimensionalized values are listed below:

$$\begin{aligned} b &= 6200 \\ \beta &= 500 \\ T_{\max} &= 3.5 \end{aligned} \quad (9)$$

Furthermore, the constants in (5) are chosen as follows:

$$\begin{aligned} A_1 &= 0.0095 \\ A_2 &= 25 \\ A_3 &= 0.953467778 \\ A_4 &= 0.036 \end{aligned} \quad (10)$$

For the numerical solution it is assumed that the rocket is initially at rest at the surface of the

earth, and that its fuel mass is 40% of the rocket total mass.

3. PROBLEM ANALYSIS

Define the state vector $\tilde{x} = \text{col}(h, v, m)$, and the co-state vector $\tilde{\lambda} = \text{col}(\lambda_h, \lambda_v, \lambda_m)$. Then the variational Hamiltonian takes the form:

$$\mathcal{H}(\tilde{\lambda}, \tilde{x}, T) = \lambda_h \dot{h} + \lambda_v \dot{v} + \lambda_m \dot{m} \quad (11)$$

where the propagation of the co-state vector obeys the equation

$$-\frac{\partial \mathcal{H}}{\partial \tilde{x}} \quad (12)$$

Using (3) and (12), and noting that the control T appears linearly in the equations of motion, one obtains for the Hamiltonian the following form:

$$\mathcal{H} = \mathcal{H}_0 + T \mathcal{H}_1 = 0 \quad (13)$$

where \mathcal{H}_0 and \mathcal{H}_1 are given by:

$$\mathcal{H}_0 = \lambda_h v + \lambda_v \left(\frac{D}{m} + h^{-2} \right) \quad (14)$$

$$\mathcal{H}_1 = \frac{\lambda_v}{m} \quad (15)$$

\mathcal{H}_1 is the "switching function" and governs the history of the control. Using Pontryagin's Maximum Principle (Pontryagin *et al.*, 1962), three possibilities exist for an extremal control, depending on the sign of the switching function:

$$\begin{aligned} T^* &= T_{\max} \quad \text{when} \quad \mathcal{H}_1 < 0 \\ 0 \leq T^* \leq T_{\max} \quad \text{when} \quad \mathcal{H}_1 &= 0 \\ T^* &= 0 \quad \text{when} \quad \mathcal{H}_1 > 0 \end{aligned} \quad (16)$$

The second case indicates the possibility of an interval of singular control, i.e. an interval of finite duration over which the \mathcal{H}_1 vanishes identically. The following relationships must then be fulfilled simultaneously on a singular arc:

$$\mathcal{H}_1 = \dot{\mathcal{H}}_1 = \ddot{\mathcal{H}}_1 = \dots = 0 \quad (17)$$

The above equations along with (13) define a manifold $E(v, m, h) = 0$ in the three-dimensional state space of v, m, h . This manifold, often called the *singular surface*, represents the locus of all possible state trajectories, corresponding to singular control effort. Note also, that $E(v, m, h) = 0$ is also the *singular control switching boundary*, since any point of the state space which does not lie on it, must feature a bang-bang control.

The three possible types of subarcs that may appear in the solution of an optimal trajectory have already been examined; however, the composite optimal trajectory consisting of these

three types of subarcs need to be determined. The analysis of the problem is complicated by the fact that the optimal solution, in general, consists of some combination of singular and non-singular subarcs, the number and sequence of which are not known *a priori*. In fact, the manner in which singular subarcs enter into composite candidates will be determined in part, by the specified two-point boundary conditions for the Euler equations. Hence, the determination of the optimal composite trajectory involves the solution of a two-point boundary value problem, frequently by means of a trial and error procedure. The next section describes a methodology that simplifies problems involving both singular and nonsingular subarcs and that can be used to determine the possible composite optimal structure.

4. TRANSFORMATION TO REDUCED STATE-SPACE

A transformation approach suggested by Kelley (1964a, b) is sometimes helpful and permits analysis of singular arcs in a state space of reduced dimension. The singular arcs become nonsingular, thus the available necessary conditions can be applied. However, this approach has the practical shortcoming that the solution of the transformation requires a closed form solution to a system of nonlinear differential equations. Fortunately, this transformation can be obtained rather easily for the present problem, allowing the structure of the problem to be studied in a reduced, two-dimensional, state-space. This is quite attractive; the complete family of singular extremals for given initial conditions can be pictured in two-dimensional space.

Omitting for brevity the theory of the transformation, Kelley *et al.* (1967) have shown, that the transformation of the original state vector \hat{x} to the canonical form leads to the new state vector \hat{z} with components

$$z_1 = h, \quad z_2 = v, \quad z_3 = mc'' \quad (18)$$

The differential equations in the new state space are derived directly from (3) and (18).

$$\begin{aligned} \dot{z}_2 &= \frac{T-D}{z_1} \exp(z_2/c) - z_1^{-2} \\ \dot{z}_3 &= -\frac{D}{c} \exp(z_2/c) - \frac{z_3}{c} z_1^{-2} \end{aligned} \quad (19)$$

The Hamiltonian for the new system is given by

$$\tilde{\mathcal{H}} = \kappa_{z_1} \dot{z}_1 + \kappa_{z_2} \dot{z}_2 + \kappa_{z_3} \dot{z}_3 \quad (20)$$

where $\tilde{\kappa} = \text{col}(\kappa_{z_1}, \kappa_{z_2}, \kappa_{z_3})$ is the co-state vector of the new state-space, satisfying the differential equations

$$\begin{aligned} \dot{\kappa}_{z_1} &= \frac{\kappa_{z_1}}{c} \left[\exp(z_2/c) \frac{\partial D}{\partial z_1} - 2z_1 z_1^{-3} \right] \\ &\quad - \kappa_{z_2} \left[-z_1^{-1} \frac{\partial D}{\partial z_1} \exp(z_2/c) + 2z_1^{-3} \right] \\ \dot{\kappa}_{z_2} &= -\kappa_{z_1} + \frac{\kappa_{z_2}}{c} \exp(z_2/c) \left[\frac{\partial D}{\partial z_2} + \frac{D}{c} \right] \\ &\quad - \kappa_{z_3} \exp(z_2/c) \frac{\partial D}{\partial z_2} + \frac{T-D}{c} \\ \dot{\kappa}_{z_3} &= \frac{\kappa_{z_3}}{c} z_1^{-2} - \kappa_{z_2} \left[-\frac{T-D}{z_1^2} \exp(z_2/c) \right] \end{aligned} \quad (21)$$

Notice from (19) that the control T appears only in one of the state equations, namely in the equation for \dot{z}_2 . One can therefore discard this equation, for analysis of the singular portion of the trajectory, and consider the z_2 variable as a new "control-like" variable, in the reduced state-space of variables z_1 and z_3 . This change occurs through the identical vanishing of the Lagrange multiplier associated with the second equation of the state. Indeed, the switching function of the transformed problem is given as

$$\tilde{\mathcal{H}}_1 = \frac{\partial \tilde{\mathcal{H}}}{\partial T} = \frac{\kappa_{z_2}}{z_1} \exp(z_2/c) \quad (22)$$

and along a singular arc we require

$$\kappa_{z_2} \neq 0 \quad (23)$$

because, throughout the trajectory,

$$\frac{\exp(z_2/c)}{z_1} \neq 0 \quad \text{always.} \quad (24)$$

The vanishing of κ_{z_2} , along the singular portion of the trajectory, can be verified through an analogous transformation for the co-state vector $\tilde{\lambda}$ of the original state space as follows: Optimal control theory indicates that the co-state variables have a special meaning: as Breakwell (1961), and Cicala (1957) have pointed out, the value at some time t of the Lagrange multiplier λ_i , associated with the variable x_i is just $\partial \mathcal{J} / \partial x_i(t)$, where \mathcal{J} is the "payoff" function with t regarded as a starting time. This interpretation of the co-states is very instructive and it will be used extensively later on. Requiring that the cost function and the Hamiltonian remain unchanged under the transformation, and using the chain rule of differentiation, the following relationship must hold along the trajectory:

$$\tilde{\lambda} = \frac{\partial \mathcal{J}}{\partial \tilde{x}} = \frac{\partial \tilde{\mathcal{H}}}{\partial \tilde{x}} \quad (25)$$

According to the interpretation mentioned earlier,

$$\dot{\bar{\kappa}} = \frac{\partial \mathcal{H}}{\partial \bar{z}} \quad (26)$$

is the co-state vector for the new state space and

$$[J] = \frac{\partial \bar{z}}{\partial \bar{x}} \quad (27)$$

is the Jacobian of the transformation, with elements

$$J_{ij} = \frac{\partial z_j}{\partial x_i} \quad i, j = 1, 2, 3. \quad (28)$$

Assuming that the transformation is nonsingular, the inverse of the Jacobian matrix exists, and the system of (25) has the unique solution

$$\bar{\kappa} = [J]^{-1} \tilde{\lambda} \quad (29)$$

which can be written analytically as

$$\begin{bmatrix} \kappa_{z_1} \\ \kappa_{z_2} \\ \kappa_{z_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -m/c \\ 0 & 0 & e^{-mz_2/c} \end{bmatrix} \begin{bmatrix} \lambda_h \\ \lambda_v \\ \lambda_m \end{bmatrix} \quad (30)$$

or in expanded form

$$\kappa_{z_1} = \lambda_h \quad (31a)$$

$$\kappa_{z_2} = \lambda_v - \frac{m}{c} \lambda_m \quad (31b)$$

$$\kappa_{z_3} = \lambda_m e^{-mz_2/c}. \quad (31c)$$

Notice that since

$$\det [J] = e^{mz_2/c} \neq 0 \quad (32)$$

the transformation is nonsingular everywhere. Equation (31b) offers another justification for the vanishing of κ_{z_2} along a singular arc. Comparing (15) and (31b) we see that the co-state κ_{z_2} is just the switching function of the original problem times the mass; consequently, along the singular arc κ_{z_2} should be identically zero. The co-state equations reduce to

$$\dot{\kappa}_{z_1} = -\frac{\partial \mathcal{H}}{\partial z_1} - \frac{\kappa_{z_3}}{c} \left[\exp(z_2/c) \frac{\partial D}{\partial z_1} - 2z_1 z_1^{-1} \right] \quad (33a)$$

$$\dot{\kappa}_{z_3} = -\frac{\partial \mathcal{H}}{\partial z_3} = \frac{\kappa_{z_3}}{c} z_1^{-2} \quad (33b)$$

and the Hamiltonian simplifies to

$$\mathcal{H} = \kappa_{z_1} z_2 - \frac{\kappa_{z_3}}{c} [D \exp(z_2/c) + z_1 z_1^{-2}]. \quad (34)$$

Notice that the new control z_2 does not appear linearly in the system of state and co-state equations, and the classical Legendre-Clebsch necessary conditions can be applied successfully in the reduced-state-space problem.

The extremals, then, of the transformed problem are the singular extremals of the original, and those extremals satisfying the strengthened version of the classical Legendre-Clebsch condition are maximizing, at least over short intervals. The stationary solution of the transformed problem corresponding to the singular subarc of the original problem occurs then, when

$$\frac{\partial \mathcal{H}}{\partial z_2} = \kappa_{z_1} - \frac{\kappa_{z_3}}{c} \exp(z_2/c) \left[\frac{D}{c} + \frac{\partial D}{\partial z_2} \right] = 0 \quad (35)$$

and the Legendre-Clebsch necessary condition requires, for a maximizing extremal

$$\frac{\partial^2 \mathcal{H}}{\partial z_2^2} = \frac{\kappa_{z_3}}{c^2} \exp(z_2/c) \left[\frac{D}{c^2} + \frac{2}{c} \frac{\partial D}{\partial z_2} + \frac{\partial^2 D}{\partial z_2^2} \right] \leq 0. \quad (36)$$

The latter relationship assures the convexity of the Hamiltonian in the neighborhood of a solution of (35), i.e. an optimal control obtained by (35) provides *at least a local* maximum of \mathcal{H} .

5. NECESSARY CONDITIONS

The fact that a trajectory satisfies the Euler differential equations and the first-order necessary conditions, only guarantees its stationary character. To determine whether a maximum is attained, further investigation is in order. Thus, the Legendre-Clebsch, Weierstrass and Jacobi conditions must be checked. Each of these three conditions is a necessary condition for a maximum. All of them, suitably strengthened, in combination with the first-order necessary conditions, provide a sufficient condition. In this section, we will briefly review the available necessary conditions for the optimality of the trajectory, in the case when singular subarcs are considered as possible candidates.

Kelley condition. The mere presence of singular members of the state-Euler system solutions does not assure the appearance of such subarcs in an optimal trajectory. In fact, as Johnson and Gibson (1963) pointed out, a singular solution may not be optimal even locally. To determine local optimality a further investigation is in order. Thus, the so-called Generalized Legendre-Clebsch, or Kelley-Contensou condition must be checked, see Kelley (1964a) and Robbins (1967). This condition can be stated as follows:

$$(-1)^q \frac{\partial}{\partial T} \left[\frac{d^{2q}}{dt^{2q}} \left(\frac{\partial \mathcal{H}}{\partial T} \right) \right] \leq 0 \quad (37)$$

where q is the order of the singularity of the arc. **Junction conditions.** An admissible control must

satisfy other requirements, in addition to satisfying the given physical constraints. If the solution is totally nonsingular, or totally singular, necessary conditions for optimality testing are available in a large number of cases. The continuity of $\mathcal{H}(t)$ and the continuity of $\tilde{\lambda}(t)$ across junctions between subarcs, the so-called *Weierstrass–Erdmann* corner condition, is perhaps the most important. However, the character of optimal trajectories which include both singular and nonsingular subarcs is less easily decided. The first results concerning the behavior of the optimal control at a junction between singular and nonsingular subarcs were derived by Kelley *et al.* (1967), and may be summarized as follows: If q is the order of the singular subarc, then

If q is *odd* a jump discontinuity in control may occur at a junction between a locally minimizing singular subarc, i.e. a subarc on which the generalized Legendre–Clebsch condition is satisfied in strengthened form, with a nonsingular subarc.

If q is *even*, jump discontinuities in control from singular subarcs satisfying the strengthened form of the generalized Legendre–Clebsch condition are ruled out.

Johnson (1967) recognized the conflict between the generalized Legendre–Clebsch condition and the junction condition for q even, and showed that analytic junctions with jumps can occur only if q is odd, but he did not identify the character of junctions between nonsingular and q even singular subarcs.

McDanell and Powers (1971), motivated by the preliminary results obtained by Kelley *et al.* (1967) and Johnson (1967), considered the problem concerning the continuity and smoothness properties of the optimal control at a junction between singular and nonsingular subarcs in more detail, and generalized the previous conclusions, with one important exception; they proved the possibility of a continuous junction for control saturation with zero slope for q odd problems, a possibility which had not been included by Kelley *et al.* and which was later ruled out for $q > 1$ by Berschanskiy (1979). Their main result was that—for analytic junctions—the sum of the order of the singular subarc and the order of the lowest time derivative of the control which is discontinuous at the junction must be an odd integer when the strengthened generalized Legendre–Clebsch condition is satisfied.

In the McDanell and Powers results, the assumption that the control is piecewise analytic is not to be taken lightly because the junction is

typically nonanalytic not only for q even, but also for q odd with $q > 1$. In fact, according to Berschanskiy, the McDannell–Powers necessary conditions are actually of interest only for $q = 1$. As was shown in his work, for q even problems or for q odd problems with $q > 1$ the transition from a nonsingular to a singular subarc is associated with *chattering* junctions, i.e. controls that switch rapidly between the upper and the lower bound faster and faster, with a point of accumulation, and which although measurable, are nonanalytic.

Jacobi and Jacobi-like conditions. Testing of the second variation, on the other hand, such as Jacobi and Jacobi-like testing is rarely carried out for nonsingular extremal candidates, and even more rarely for candidates with isolated singular points, possibly corners, as pointed out by Kelley and Moyer (1985). Extremals corresponding to the second case, so-called broken extremals, have been studied with generality, detail and rigor by Larew (1919), Reid (1935) and Caratheodory (1967). Moyer (1965, 1970) using this idea, developed a computational technique in the case of a nonsingular extremal exhibiting corners, and used this approach successfully in an orbital transfer. However, Jacobi-like testing for composite Euler solutions including singular subarcs is still a research area, and the few attempts made in this direction, mainly due to McDanell and Powers (1970), are limited to the case of a totally singular arc. Very few methods have been also developed for the more complex case of a composite extremal, mainly by Speyer and Jacobson (1971a, b) and Moyer (1973).

All the above conditions, though only necessary, help to eliminate some of the possible subarc-sequence candidates.

6 CONTROL-LOGIC ANALYSIS

Only the free-time case was studied, but the method of solution is applicable also for any value of fixed final time. Due to the sharp increase of the partial $\partial C_D / \partial M$ near Mach 1, the singular surface witnesses also a peak in the same region (Fig. 2). Moreover, projections of the singular surface into mass-velocity and altitude-velocity planes reveal the existence of a nonadmissible portion of the variable thrust arc, since it corresponds to increasing mass (shown by a dashed line in Figs 3 and 4). Therefore, an optimal switching structure cannot include a singular arc in the transonic region, on account of the violation of the requirement the mass be monotonically nonincreasing.

The problem becomes more transparent if

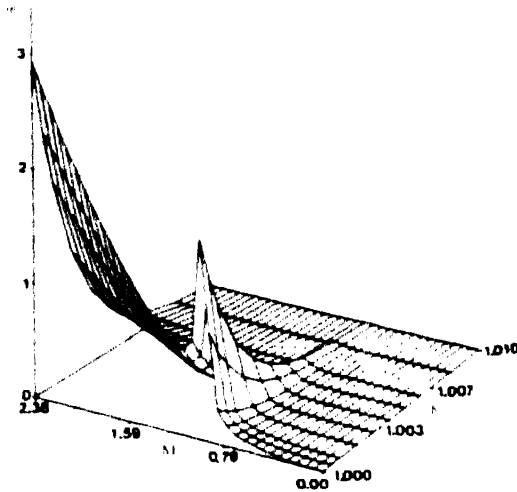


FIG. 2. Singular surface $E(v, m, h) = 0$ for Mach dependent drag coefficient

one uses the transformation to z_1 and z_2 state-space described before. It should be noted however, that such an approach is equivalent to admitting jump discontinuities in the new control variable $z_2 = v$. Such discontinuities, occurring at corner points of the solution, imply impulsive behavior of the thrust T . Such impulsive behavior would be admissible in the absence of inequality constraints on T , but in practice, there is always a limit on the available thrust output. However, thrust impulses, while not physically possible, are convenient idealizations to very rapid burning of fuel. Thus, an optimal solution obtained in the z -space would still be of importance as an approximation to the case of a very high magnitude of the throttle setting, and in addition to this, it could provide physical insight to the problem.

The analysis can be stated briefly as follows: examine the singular arc by transforming to

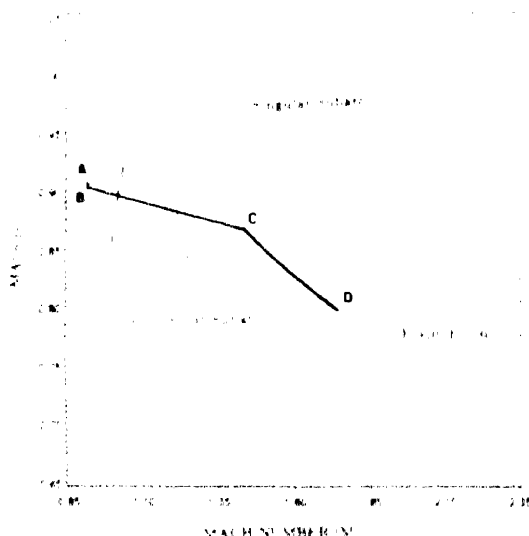


FIG. 3. Projection of the trajectory on the mass-velocity plane for $T_{max} = \infty$

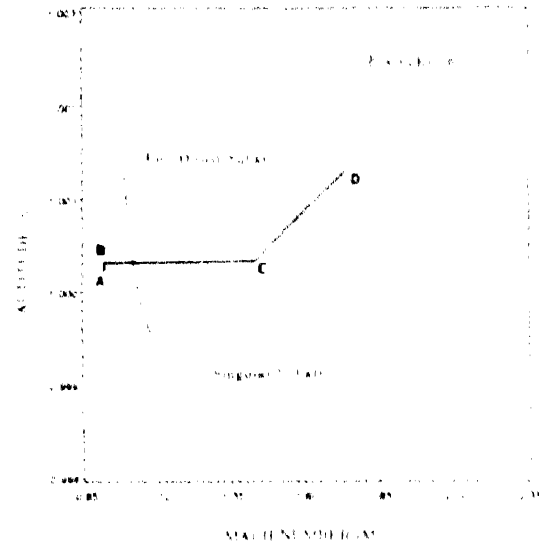


FIG. 4. Projection of the trajectory on the altitude-velocity plane for $T_{max} = \infty$

z -space with a new control-like variable, z_2 , which maximizes the new Hamiltonian. The variation of the Hamiltonian vs the velocity along the extremal, corresponding to the singular arc of the original problem, reveals that along this singular arc the Hamiltonian has three stationary values, corresponding to three solutions of the equation of the singular surface $E(v, m, h) = 0$. Two of those correspond to a maximum, and the other corresponds to a minimum value of the Hamiltonian. The first maximum corresponds to the subsonic branch, the minimum corresponds to the transonic branch, screened out, and the second maximum corresponds to the supersonic branch of the singular surface. Henceforth we shall use the terms "subsonic maximum" or "subsonic solution", and "supersonic maximum" or "supersonic solution" to distinguish between the two cases of interest. Thus, points corresponding to the transonic solution cannot be included in an optimal trajectory for a second reason, since such points provide a local minimum rather than a maximum for the Hamiltonian.

From Fig. 5 we notice that there is a point in time t_{tr} when both solutions provide the same maximum to the Hamiltonian, and the velocity then *jumps* from the subsonic to the supersonic solution. That is,

$$\mathcal{H}(v_{sub}(t_{tr})) = \mathcal{H}(v_{sup}(t_{tr})) \quad (38)$$

where the subscript "sub" denotes the subsonic solution, and the subscript "sup" denotes the supersonic solution. Hence,

$$z_2(t) = v_{sub} \quad \text{for } t \leq t_{tr} \quad (39a)$$

and

$$z_2(t) = v_{sup} \quad \text{for } t \geq t_{tr}. \quad (39b)$$

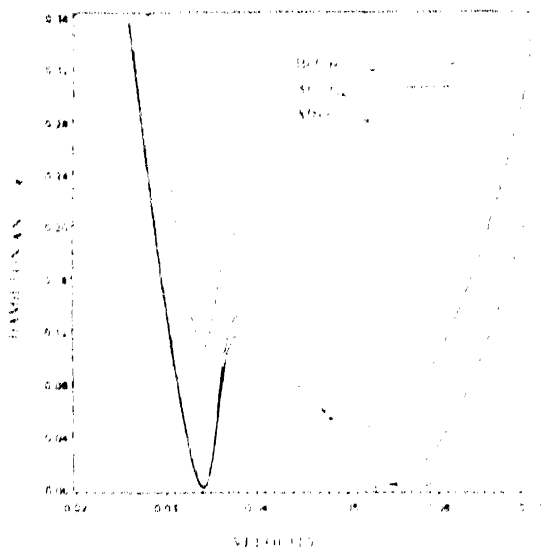


FIG. 5. Variation of H with v , $z_2 = v$

That is because, although both the subsonic and the supersonic solution give a relative maximum, an optimal control should correspond to the absolute maximum of the Hamiltonian. However, the result is limited to the case in which no upper bound on thrust is imposed.

7. COMPOSITE OPTIMAL TRAJECTORY

The previous analysis indicates that an optimal trajectory should start with a full-thrust subarc until the subsonic solution of the singular surface is reached. Then a variable-thrust subarc using this solution is used up to the point when both the subsonic and the supersonic branches provide the same maximum to the Hamiltonian. A switching then to the supersonic branch occurs, and the trajectory remains on the singular surface until the time when the fuel is exhausted. Then a final coasting arc is used, until the terminal boundary conditions are satisfied.

Although this thrust history would provide the optimal switching structure for the case of $T_{max} = \infty$, this will not be necessarily true for the case of bounded thrust. In such a case discontinuities in the velocity are of course unacceptable, and the validity of the solution depends on the value of the upper bound of the thrust. Thus, the structure of the optimal trajectory is still in question. This is the topic of the following section.

8. BOUNDED-THRUST CASE

The analysis so far shows that the variational problem has a special mathematical structure, in so far as the occurrence of two optimal solutions of $E(v, m, h) = 0$ implies the existence of an infinite number of composite solutions, in the passing through the transonic region, all

satisfying the same boundary conditions. The question is: which of this family extremals is to be preferred from the point of view of maximizing the altitude? For the case of unbounded thrust the answer has already been given: At a time $t = t_{sw}$, when the Hamiltonian in the reduced z -space switches from its subsonic to its supersonic maximum. Although valid only for unbounded thrust, nevertheless, this remark gives us a hint: an optimal trajectory must *accelerate* from the subsonic to the supersonic region. Since the variable-thrust case must be ruled out, our only choice is the use of full thrust between the two solutions of $E(v, m, h) = 0$. Furthermore, because for a realistic case $T_{max} < \infty$, the switching from the variable-thrust to the second full-thrust subarc must take place somewhere *before* the time t_{sw} , and such that the switching function vanishes at the points of departure and arrival to the singular surface (points B and D in Figs 3 and 4). In addition to this, the switching function should remain positive all along the full-thrust subarc in order to satisfy the optimality condition of (16).

Thus, a trial-and-error procedure is needed to determine the points B and D. The result obtained, using the boundary-problem solver BOUNDSOL (Bulirsch, 1971), was rather disappointing; an optimal switching from the first variable-thrust subarc to the second full-thrust subarc (point B), should take place before the switching of the first full-thrust subarc to the first variable-thrust subarc (point A). Therefore, for the case of $T_{max} = 3.5$, an optimal trajectory cannot have this switching structure, but rather must have the simpler full-singular-coast sequence, with the singular subarc corresponding to the supersonic solution of $E(v, m, h) = 0$.

However, when an analogous calculation for the case $T_{max} = 6$ was performed, the new, more complex, sequence of subarcs full-singular-full-singular-coast, gave indeed a *higher* final altitude than the full-singular-coast sequence (Table 1). In Figs 6-8 is shown the history of the three components of the nondimensionalized state vector $\vec{x} = (h, v, m)$ respectively, for the optimum burning program corresponding to $T_{max} = 6$. The switching sequence for this optimum thrust program is depicted in Fig. 9. Notice the

TABLE 1. COMPARISON BETWEEN THE CLASSICAL FULL-SINGULAR-COAST SEQUENCE AND THE NEW FULL-SINGULAR-FULL-SINGULAR-COAST SEQUENCE INVOLVING A FULL-THRUST SUBARC AT THE TRANSONIC REGION, FOR THE CASE OF $T_{max} = 6$

	Final time	Final altitude
F-S-C	0.197374	1.0132976
F-S-F-S-C	0.198978	1.0133038

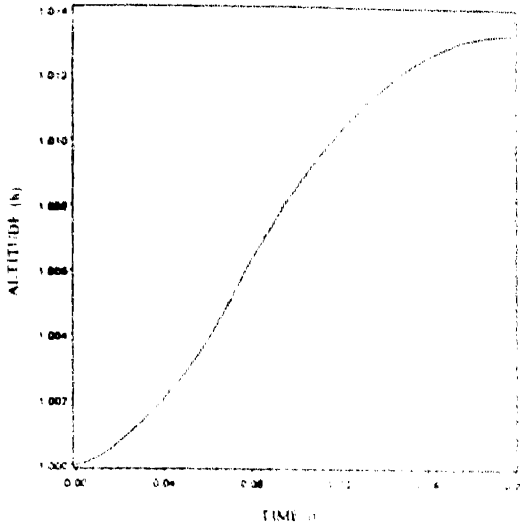


FIG. 6. Variation of altitude h with time for $T_{\max} = 6$

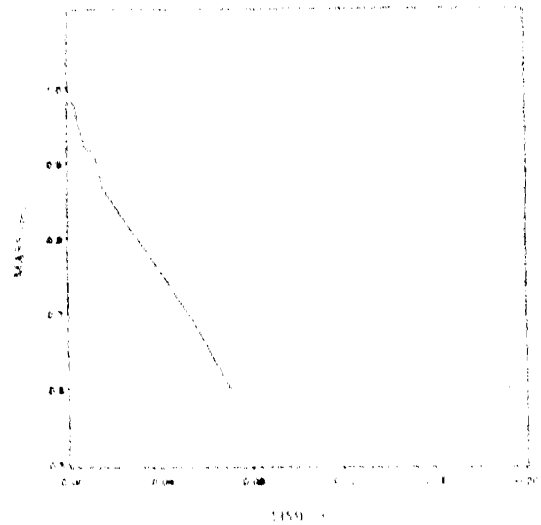


FIG. 8. Variation of mass m with time for $T_{\max} = 6$

corner points during the velocity and mass evolution, that correspond to the points of discontinuity of the thrust. This is the result of the control variable T entering directly to the right-hand side of the equations of motion for the velocity and mass, (3). On the other hand, the evolution of the state variable representing the altitude is smooth, since the thrust does not appear to the corresponding state equation.

10. CONCLUSIONS

The problem of maximizing the final altitude of a vertically ascending rocket has been analyzed for the case of bounded thrust, and quadratic drag law, with the drag coefficient as a function of the Mach number, witnessing a sharp increase in the transonic region. A more complex switching structure, with an intermediate full-thrust subarc in transition through the transonic region, was required owing to the

requirement that the mass should be monotonically nonincreasing. The results are identical with those of Garfinkel, for the $T_{\max} = \infty$ case, although a totally different approach was used. The solution, using a transformation to a reduced two-dimensional state space, showed that the optimality of the solution depends on the assumed upper bound on the thrust. Numerical results obtained verified the superior performance of the new thrust program, over the classical full-singular-coast sequence, at least for a sufficiently high upper bound on the thrust.

The Kelley necessary condition for singular arcs, and the McDaniel and Powers condition for joining singular and nonsingular subarcs were checked, and were found to be satisfied.

A companion paper by the authors examines time-of-flight constraint effects in the problem (Tsiotras and Kelley, 1988).

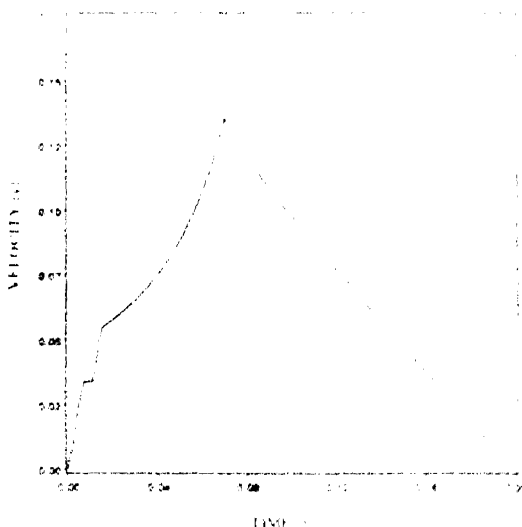


FIG. 7. Variation of velocity v with time for $T_{\max} = 6$

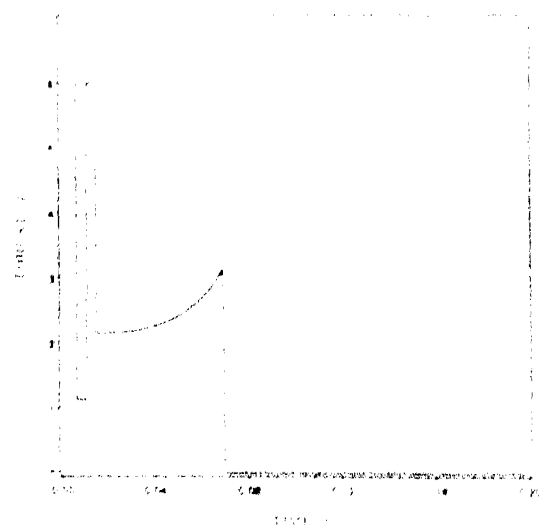


FIG. 9. Variation of thrust T with time for $T_{\max} = 6$

Acknowledgements—The authors would like to thank Dr E. M. Cliff for his valuable suggestions during the preparation of this work. Support for this research was provided by the USAF Armament Laboratory, Eglin AFB, Florida under contract F08635-86-K-0390

REFERENCES

- Bershtanskii, Y. M. (1979) Conjugation of singular and nonsingular parts of optimal control. *Aut. Remote Control*, **40**, 325–330.
- Breakwell, J. V. (1961) Approximations in flight optimization techniques. *Aerospace Engng*, **20**, 26–27.
- Bryson, A. Jr. and S. E. Ross (1958) Optimum rocket trajectories with aerodynamic drag. *Jet Propulsion*, **28**, 465–469.
- Bulirsch, R. (1971) Einführung in die Flugbahnoptimierung die Mehrzielmethode zur Numerischen Lösung von Nichtlinearen Randwertproblemen und Aufgaben der Optimalen Steuerung. *Lehrgang Flugbahnoptimierung*, Carl-Drans-Gesellschaft e.v.
- Caratheodory, C. (1967) *Calculus of Variations and Partial Differential Equations of the First Order*, Vol. 2. Holden-Day, San Francisco.
- Cicala, P. (1957) *An Engineering Approach to the Calculus of Variations*. Levrotto e Bella, Turin.
- Gärtinkel, B. (1963) A solution of the Goddard problem. *SIAM J. Control*, **1**, 349–368.
- Goddard, R. H. (1919) *A Method of Reaching Extreme Altitudes*. Smithsonian Institution Misc. collections, **71** (reprint by American Rocket Society, 1946).
- Hamel, G. (1927) Über eine mit dem Problem der Rakete Zusammenhängende Aufgabe der Variationsrechnung. *Zeitschrift für angewandte Mathematik und Mechanik*, **7**, 451–452.
- Johnson, C. D. and J. E. Gibson (1963) Singular solutions in problems of optimal control. *IEEE Trans. Aut. Control*, **8**, 4–15.
- Johnson, C. D. (1967) Singular solutions in problems of optimal control. In Leondes, C. T. (ed.) *Adv. Control Syst.*, **2**, Academic Press, New York.
- Kelley, H. J. (1964a) A transformation approach to singular subarcs in optimal trajectory and control problems. *SIAM J. Control*, **2**, 234–240.
- Kelley, H. J. (1964b) A second variation test for singular extremals. *AIAA J.*, **2**, 1380–1382.
- Kelley, H. J., R. E. Kopp and H. G. Moyer (1967) Singular extremals. In Leitmann, G. (ed.), *Topics in Optimization*. Academic Press, New York.
- Kelley, H. J. and H. G. Moyer (1985) Computational Jacobi-test procedure. *Current Trends Control*. Proceedings of workshop. Dubrovnik, Yugoslavia, JUREMA, Zagreb, 1985.
- Larow, G. A. (1919) Necessary conditions in the problem of Mayer in the calculus of variations. *Trans. Am. Math. Soc.*, **20**, 1–22.
- Leitmann, G. (1956) A calculus of variations solution of Goddard's problem. *Astronaut. Acta*, **2**, 55–62.
- Leitmann, G. (1957) Optimum thrust programming for high altitude rockets. *Aeronaut. Engng Rev.*, **16**, 63–66.
- Leitmann, G. (1963) An elementary derivation of the optimal control conditions. *12th Int. Astronaut. Congr.*, Academic Press, New York.
- McDanell, J. P. and W. F. Powers (1970) New Jacobi-type necessary and sufficient conditions for singular optimization problems. *AIAA J.*, **8**, 1416–1420.
- McDanell, J. P. and W. F. Powers (1971) Necessary conditions for joining optimal singular and nonsingular subarcs. *SIAM J. Control*, **9**, 161–173.
- Miele, A. (1955) Optimum climbing technique for a rocket-powered aircraft. *Jet Propulsion*, **25**, 385–391.
- Miele, A. and P. Cicala (1956) Generalized theory of the optimum thrust programming for the level flight of a rocket-powered aircraft. *Jet Propulsion*, **26**, 443–455.
- Miele, A. and C. R. Cavoti (1958) Generalized variational approach to the optimum thrust programming for the vertical flight of a rocket. *ZFW*, **6**, 102–109.
- Miele, A. (1962) Extremization of linear integrals Green's theorem. *Optimization Techniques with Applications to Aerospace Systems*. Academic Press, New York, 69–98.
- Moyer, H. G. (1965) Minimum impulse coplanar circle-ellipse transfer. *J. AIAA*, **3**, 209–267.
- Moyer, H. G. (1970) Optimal control problems that test for envelope contacts. *J. Optimiz. Theory Applic.*, **6**, 287–298.
- Moyer, H. G. (1973) Sufficient conditions for a strong minimum in singular control problems. *SIAM J. Control*, **11**, 620–636.
- Pontryagin, L. S., V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mishchenko (1962) *The Mathematical Theory of Optimal Processes*. Interscience, New York.
- Reid, W. T. (1935) Discontinuous solutions in the nonparametric problem of Mayer in the Calculus of variations. *Amer. J. Math.*, **57**, 69–93.
- Robbins, H. M. (1967) A generalized Legendre-Clebsch condition for the singular cases of optimal control. *IBM J. Res.*, **11**, 361–372.
- Speyer, J. L. and D. H. Jacobson (1971a) Necessary and sufficient conditions for optimality for singular control problems: A transformation approach. *J. Math. Anal. Applic.*, **33**, 163–187.
- Speyer, J. L. and Jacobson, D. H. (1971b) Necessary and sufficient conditions for optimality for singular control problems: A limit approach. *J. Math. Anal. Applic.*, **34**, 239–266.
- Tsien, H. S. and R. C. Evans (1951) Optimum thrust programming for a sounding rocket. *J. Amer. Rocket Soc.*, **21**, 99–107.
- Tsotras, P. and H. Kelley (1988) The Goddard problem with constrained time-of-flight. *Proc. Am. Control Conf.*, Atlanta, 1412–1421, also *AIAA J. Guidance, Control Dynam.* (to appear).
- Zlatskiy, V. T. and B. N. Kiforenko (1983) Computation of optimal trajectories with singular-control sections. *Vychislitel'naya i Prikladnaya Matematika*, **49**, 101–108.

Vibrational Control of Nonlinear Time Lag Systems: Vibrational Stabilization and Transient Behavior*

JOSEPH BENTSMAN,^{†‡} KEUM S. HONG[†] and JAMEL FAKHFAKH[†]

Periodic excitations introduced into time lag systems can stabilize unstable equilibria or induce stable periodic solutions with the desired properties. The resulting open-loop technique termed vibrational control can be useful when on-line measurements are not available.

Key Words—Nonlinear systems; time lag systems; stability

Abstract—This paper addresses two problems of control of nonlinear time lag systems: (i) an existence and a synthesis of parametric vibrations for their stabilization and (ii) the transient behavior analysis of the vibrationally controlled nonlinear systems with time lags. In this work, stabilizability conditions for two vibration types are formulated and procedures for the synthesis of the corresponding stabilizing vibrations are proposed. The method for the transient behavior analysis of vibrationally controlled systems on a finite time interval is developed as well. Several examples are given to support the theory presented.

1. INTRODUCTION

VIBRATIONAL CONTROL is an open-loop technique that utilizes parametric excitation of a dynamical system for achieving control objectives. A well-known example of the vibrational control effect is a stabilization of an inverted pendulum by vertical oscillations of its support. Obviously, in this case there is no interference into the plant structure, and the control objective to keep the pendulum in the upright position is achieved by much simpler means than using feedback. An extensive theoretical and experimental comparison of vibrational control with feedback and feedforward control strategies is given by Meerkov (1980), Cinar *et al.* (1987), Bentsman

and Hvostov (1988) and Fakhfakh and Bentsman (1990). These studies show that being an open-loop technique, vibrational control can (1) stabilize the plants when on-line measurements and hence feedback, are impossible, such as in powerful continuous CO₂ lasers and particle accelerators; or (2) under the practical restrictions on sensing and actuation create desired stable operating regimes unattainable by feedback for such plants as chemical reactors and laser illuminated reactions. The mathematical machinery of vibrational control has also found important applications in the synthesis of linear periodic feedback controllers (Lee *et al.*, 1987) that ensure an infinite gain margin in the robust stabilization of the nonminimum phase plants with the right half plane poles, which is not possible with a linear time invariant feedback (cf. Khargonekar *et al.*, 1985).

The theory of vibrational control for systems governed by linear and nonlinear ordinary differential equations has been developed by Meerkov (1980), and Bellman *et al.* (1986a, b) and Bentsman (1987), respectively. However, many physical systems with nonlinear behavior such as chemical reactions and combustion processes have time delayed states (cf. Ray 1981; Kolmanovskii and Nosov, 1986). This motivates studies of oscillatory stabilizing effects in nonlinear systems with time lags.

This paper introduces the first results in vibrational control of nonlinear systems with time delays: conditions for the existence of stabilizing vibrations and a procedure for their synthesis are given for a class of nonlinear time lag systems, a method for the transient behavior analysis of vibrationally controlled systems is proposed, and examples of vibrational stabi-

* Received 23 January 1989, revised 27 June 1990, received in final form 26 July 1990. The original version of this paper was presented at the IFAC Symposium on *Nonlinear Control System Design* which was held in Capri, Italy during June, 1989. The published proceedings of this IFAC Meeting may be ordered from Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor V. Utkin under the direction of Editor H. Kwakernaak.

[†] Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, 1206 West Green Street, Urbana, IL 61801, U.S.A.

[‡] Author to whom all correspondence should be addressed.

lization of nonlinear time lag systems are presented.

In this paper a class of nonlinear systems with a finite number of constant delays is considered which is described by the equation

$$\dot{x}(t) = \sum_{i=1}^m P_i(x(t), x(t-d_i), \lambda),$$

$$P_i: R^n \times R^n \times R^l \rightarrow R^n, \quad \dot{x}(t) \triangleq \frac{dx}{dt}, \\ i = 1, \dots, m,$$

$$P_i(x(t), x(t-d_i), \lambda) \\ = [p_{i1}(x(t), x(t-d_i), \lambda), \dots, \\ p_{in}(x(t), x(t-d_i), \lambda)]^T,$$

$$P_i(0, 0, \lambda) = 0; \quad i = 1, \dots, m, \quad (1)$$

where $x \in R^n$ is a state, $\lambda = [\lambda_1, \dots, \lambda_l]^T$ are parameters subjected to vibrations, t is dimensionless time, and d_i , $i = 1, \dots, m$, are time lags of the order $O(\varepsilon)$, $0 < \varepsilon \ll 1$. The paper follows the terminology of Bellman *et al.* (1986a, b).

Introduce into (1) parametric vibrations according to the law

$$\lambda(t) = \lambda_0 + f(t) \quad (2)$$

where λ_0 is a constant vector and $f(t)$ is a periodic zero average (PAZ) vector. Then, (1) takes the form

$$\dot{x}(t) = \sum_{i=1}^m P_i(x(t), x(t-d_i), \lambda_0 + f(t)). \quad (3)$$

Throughout the paper it will be assumed that (3) can be represented as

$$\dot{x}(t) = \sum_{i=1}^m P_i(x(t), x(t-d_i), \lambda_0) + Q(f(t), x(t)) \quad (4)$$

where $Q(\cdot, \cdot)$ is a vector function linear with respect to its first argument.

Following Bellman *et al.* (1986a, b), if $Q(f(t), x(t)) = l(t)$, where $l(t)$ is a PAZ vector, the introduced vibrations are referred to as *vector additive*, if $Q(f(t), x(t)) = D(t)x(t)$, where $D(t)$ is an $n \times n$ PAZ matrix, the vibrations are called *linear multiplicative*, and if $Q(f(t), x(t)) = D(t)X(x(t))$, where $X: R^n \rightarrow R^n$ is a nonlinear map, the vibrations are termed *nonlinear multiplicative*. In the present paper, we consider vibrational stabilization and transient behavior of a class of nonlinear systems (4) with time delays d_i , $i = 1, \dots, m$, of the same order of magnitude as the period of vibrations and with linear multiplicative and vector additive vibrations. The proofs of all formal statements are given in the Appendix.

II. VIBRATIONAL STABILIZATION

Assume that (1) has an equilibrium point $x_s(t) = x_s = \text{const.}$ for a fixed λ_0 (note that $x_s(t) = x_s(t-d_i) = x_s$).

Definition 1. An equilibrium point x_s of (1) is said to be *vibrationally stabilizable* (*v-stabilizable*) if for any $\delta > 0$ there exists a PAZ vector $f(t)$ such that (3) has an asymptotically stable almost periodic solution $x^*(t)$, $-\infty < t < \infty$, characterized by

$$\|x^* - x_s\| < \delta, \quad \bar{x}^* = \overline{x^*(t)} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x^*(t) dt. \quad (5)$$

Definition 2. An equilibrium point x_s of (1) is said to be *totally vibrationally stabilizable* (*t-stabilizable*) if it is *v-stabilizable* and in addition $x^*(t) = \text{const.} = x_s$, $-\infty < t < \infty$.

The problem of vibrational stabilization consists of: (1) Finding the conditions for the existence of stabilizing vibrations (*v-* and *t-stabilizability*) and (2) finding the actual parameters of vibrations that ensure the desired stabilization.

In order to address this problem for zero equilibrium of system (1) introduce system

$$\dot{x}(t) = \sum_{i=1}^m P_i(x(t), x(t-\varepsilon r_i), \lambda) \quad (6)$$

which is obtained from (1) by replacing d_i by εr_i , $r_i = O(1)$, $i = 1, \dots, m$. In the discussion below, vibrational stabilization of (6) will be first considered and then related to that of (1).

A. Linear multiplicative vibrations

Define

$$A \triangleq \sum_{i=1}^m \frac{\partial P_i(\xi, \eta, \lambda_0)}{\partial \xi} \bigg|_{\xi = \eta = x_s = 0} \\ \text{and} \\ B_i \triangleq \frac{\partial P_i(\xi, \eta, \lambda_0)}{\partial \eta} \bigg|_{\xi = \eta = x_s = 0} \quad (7)$$

Denote by $\Phi(t)$ a fundamental matrix solution of the equation

$$\dot{x}(t) = F(t)x(t) \quad (8)$$

and introduce an ordinary differential equation

$$\dot{z}(t) = Rz(t) \quad (9)$$

where R is defined as follows:

$$R \triangleq \bar{A} + \sum_{i=1}^m \bar{B}_i, \quad (10)$$

and \bar{A} and \bar{B}_i are computed as

$$\bar{A} = \Phi^{-1}(t)A\Phi(t) \\ \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \Phi^{-1}(t)A\Phi(t) dt, \quad (11)$$

and

$$\bar{B}_i = \Phi^{-1}(t)B_i\Phi(t - \tau_i), \quad i = 1, \dots, m. \quad (12)$$

Theorem 1. Assume that there exists a sufficiently large set $\Omega \subset R^n$, $0 \in \Omega$, such that $P_i(\xi, \eta, \lambda_0)$ introduced in (1) is continuously differentiable for all $\xi, \eta \in \Omega$; $\forall i = 1, \dots, m$; then

(i) The zero equilibrium $x_i = 0$ of (6) is t -stabilizable if there exists a PAZ matrix $F(t)$ such that a fundamental matrix $\Phi(t)$, $t \in (-\infty, \infty)$, of $\dot{x} = F(t)x$ is almost periodic, and R defined in (10) is a Hurwitz matrix;

(ii) There exists positive $\varepsilon_0 = \text{const.}$ such that $x_i = 0$ of (6) is t -stabilizable by linear multiplicative vibrations $D(t)x(t) = (1/\varepsilon)F(t/\varepsilon)x(t)$, $0 < \varepsilon \leq \varepsilon_0$, where $F(t)$ satisfies all the conditions of assertion (i) above.

Corollary 1. Let an assumption of Theorem 1 hold for system (1) with fixed delays d_i , $i = 1, \dots, m$. Suppose that there exists a set of constants r_i , $i = 1, \dots, m$, such that $d_i/r_i = d_j/r_j$, $i, j = 1, \dots, m$, and $x_i = 0$ of system (6) with these constants r_i , $i = 1, \dots, m$, is t -stabilizable by linear multiplicative vibrations $(1/\varepsilon)F(t/\varepsilon)x(t)$, $0 < \varepsilon \leq \varepsilon_0$. Then the zero equilibrium of (1) is t -stabilizable by linear multiplicative vibrations $(1/\varepsilon_1)F(t/\varepsilon_1)x(t)$, $\varepsilon_1 \triangleq d_i/r_i$, if $\varepsilon_1 \leq \varepsilon_0$.

Remark 1. The significance of Corollary 1 is in that it relates the vibrational stabilization of system (1) with fixed delays d_i to the vibrational stabilization of system obtained from (1) by replacing delays d_i by quantities εr_i , $d_i/r_i = d_j/r_j$, $i, j = 1, \dots, m$, with fixed r_i s and varying ε . This leads to a constructive method for the synthesis of the stabilizing vibrations, since for fixed r_i , $i = 1, \dots, m$, Theorem 1 decomposes this synthesis into a two-stage procedure. First, PAZ matrix $F(t)$ is sought which makes R Hurwitz.

Once the desired matrix $F(t)$ is found, the second stage consists of a one-parameter (ε) computer search where equation (4) with linear multiplicative vibrations $D(t)x(t) = (1/\varepsilon)F(t/\varepsilon)x(t)$ and $d_i \rightarrow \varepsilon r_i$, $i = 1, \dots, m$, $d_i/r_i = d_j/r_j$, $\forall i, j = 1, \dots, m$, is simulated until ε_0 is found for which stability is achieved. Such ε_0 will necessarily exist for fixed r_i , $i = 1, \dots, m$, due to assertion (ii) of Theorem 1. If $d_i/r_i \neq \varepsilon_1 \leq \varepsilon_0$, then the stabilizing vibrations for (4)

with the original delays d_i , $i = 1, \dots, m$, are given by $D(t)x(t) = (1/\varepsilon_1)F(t/\varepsilon_1)x(t)$.

Remark 2. Analytical estimates of an upper bound on ε_0 are usually extremely conservative (cf. Bellman *et al.*, 1985), therefore the value of ε_0 for a given nonlinear time lag system of the form (6) is best determined by a numerical simulation as described in the previous remark.

Remark 3. The choice of a fundamental matrix $\Phi(t)$ in (8) is immaterial and is dictated only by convenience. Indeed, since any fundamental matrix $\Phi(t)$ of (8) is related to any other fundamental matrix of (8), say $\Phi_1(t)$, via a constant nonsingular matrix (denote it as C) as $\Phi(t) = \Phi_1(t)C$, from (10) by direct substitution we have

$$R = C^{-1}R_1C, \quad (13)$$

where

$$R_1 = \Phi_1^{-1}(t)A\Phi_1(t) + \sum_{i=1}^m \Phi_1^{-1}(t)B_i\Phi_1(t - \tau_i).$$

Thus, R_1 is computed exactly as in (10) but with $\Phi(t)$ replaced by $\Phi_1(t)$, and due to (13) R and R_1 have identical spectra.

Remark 4. The assumption of almost periodicity of $\Phi(t)$ in t guarantees the existence of the averages (11) and (12). The elimination of this assumption results in admitting zero mean matrices $F(t)$ for which $\Phi(t)$ is unbounded. In this case (11) and (12) do not exist. One could potentially use Lyapunov's substitution to transform $\dot{x} = F(t)x$ into a time invariant system, find Floquet multipliers, and reject matrices $F(t)$ that give rise to unbounded $\Phi(t)$. However, currently there is no constructive method for finding closed form Lyapunov's substitutions for a general class of systems $\dot{x} = F(t)x$ with periodic zero mean $F(t)$.

Remark 5. Theorem 1 and Corollary 1 show that as in the case of nonlinear systems with no delays (cf. Bellman *et al.*, 1986a), the conditions of t -stabilizability of zero equilibrium of (6) and hence of (1) by linear multiplicative vibrations depend only on the properties of the linearization of (6) or of (1) at zero. However, for other types of vibrations, this is not true as will be shown further in the paper.

Remark 6. Theorem 1 also provides a clue to the robustness of t -stabilizability properties with respect to small delays. It is seen that if the delays are of the same order of magnitude as the period of oscillations, they cannot be neglected, except for the special system structures, since

quantities $r_i = d_i/\varepsilon = O(1)$ significantly affect the spectrum of matrix R , as can be seen from its definition (10).

An example that demonstrates stabilization of a nonlinear time delay system by linear multiplicative vibrations is given next.

Example 1. Duffing equation with time lags. Consider a scalar Duffing equation with time lag $d = \text{const}$ in the states

$$\begin{aligned} \ddot{x}(t) + a_1 \dot{x}(t) + a_2 \dot{x}(t-d) - b_1 x(t) - b_2 x(t-d) \\ + c_1 x'(t) + c_2 x'(t-d) = 0, \\ a_1, a_2, b_1, b_2, c_1, c_2 > 0, \end{aligned} \quad (14)$$

which in the state space form is given by

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= -a_1 x_2(t) - a_2 x_2(t-d) + b_1 x_1(t) \\ &\quad + b_2 x_1(t-d) - c_1 x_1'(t) - c_2 x_1'(t-d), \\ x_1 &\triangleq x, \quad x_2 \triangleq \dot{x}. \end{aligned} \quad (15)$$

Equation (14) can serve as an approximate model of an inverted pendulum with a cavity filled with viscous liquid. The term with the delayed first derivative describes dissipation of the mechanical energy in the system due to viscous liquid. This term is a lumped approximation of a more complicated description of this dissipation by the retarded integrodifferential term given by Strizak (1982, p. 238). The linearization of (15) at $x_i = 0$ has the form

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= -a_1 x_2(t) - a_2 x_2(t-d) + b_1 x_1(t) \\ &\quad + b_2 x_1(t-d). \end{aligned} \quad (16)$$

Introducing vibrations $f(t) = (\alpha/\varepsilon) \cos(t/\varepsilon)$ into coefficient b_1 ,

$$b_1 \rightarrow b_1 + \frac{\alpha}{\varepsilon} \cos(t/\varepsilon), \quad (17)$$

equation (8) is given by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \alpha \cos t & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (18)$$

and its fundamental matrix is

$$\Phi(t) = \begin{bmatrix} 1 & 0 \\ \alpha \sin t & 1 \end{bmatrix}. \quad (19)$$

Replacing in (14) d by εr and computing R of (10) we obtain a corresponding averaged equation (9) given by

$$\ddot{z}(t) + r_1 \dot{z}(t) + r_2 z(t) = 0 \quad (20)$$

where

$$r_1 = a_1 + a_2, \quad r_2 = -(b_1 + b_2) + \frac{\alpha^2}{2}.$$

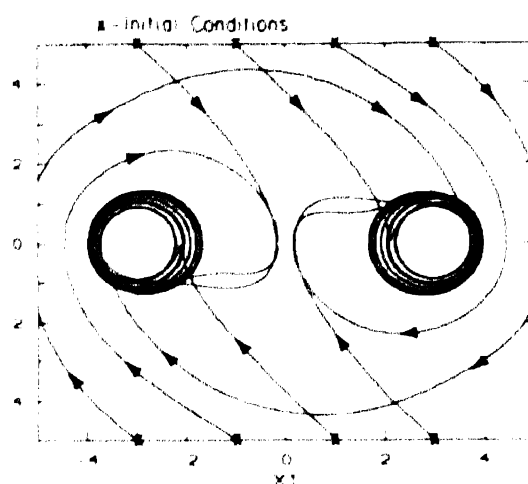


FIG. 1. Phase portrait of Duffing equation with a time lag $d = 0.4\pi$.

Therefore equation (15) with $d \rightarrow \varepsilon r$, $r = O(1)$, and vibrations introduced into the coefficient b_1 as in (17) has a stable zero equilibrium point when $a_1 + a_2 > 0$ and $\alpha > [2(b_1 + b_2)]^{1/2}$ for any positive ε smaller than some ε_0 , the existence of which for fixed r is guaranteed by Theorem 1. Consider the case when $a_1 = 0.6$, $a_2 = 0.4$, $b_1 = b_2 = 0.5$, and $c_1 = c_2 = 0.05$. Then for $\alpha > \sqrt{2}$, $r = \text{const.} = O(1)$, and $0 < \varepsilon \leq \varepsilon_0$, vibrations (17) must stabilize the originally unstable zero equilibrium point of (15) with any $d \leq \varepsilon_0 r$. Such stabilization is indeed demonstrated in Figs 1 and 2. Figure 1 shows the phase portrait of Duffing equation with time lags without parametric excitation. It is seen from this diagram that $x_i = 0$ is an unstable equilibrium point. Two other equilibrium points are given on the graph by $x_i = \pm\sqrt{10}$. Figure 2 shows a trajectory of Duffing equation with oscillations with $x_1(t) = 3$ and $x_2(t) = 5$ for $t \geq 0$, $\alpha = 2$, $d = 0.4\pi$, and $\varepsilon_1 = 0.05$. Since the trajectory in Fig. 2 converges to zero, we can

• Initial Condition

$\varepsilon = 0.05$
Time = 10 sec

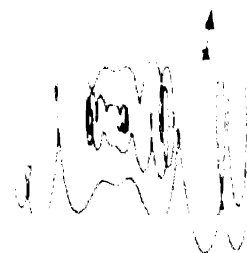


FIG. 2. A trajectory of Duffing equation with a time lag $d = 0.4\pi$ and oscillations.

observe the qualitative changes in system behavior, and specifically, a t -stabilization of an unstable zero equilibrium.

B. Vector additive vibrations

Consider the Taylor expansion of $P_i(\xi, \eta, \lambda) = P_i(\xi, \eta)$ around an equilibrium point $x_i = \xi = \eta = 0$

$$P_i(\xi, \eta) = \sum_{j=1}^{\infty} \frac{1}{j!} v_i^{(j)} + \sum_{k=1}^{\infty} \frac{1}{k!} w_i^{(k)} \quad (21)$$

where

$$v_i^{(j)} = \{[v_{i1}, \dots, v_{in}]^T\}^{(j)} \triangleq [v'_{i1}, \dots, v'_{in}]^T, \\ v'_{il} \triangleq \left(\xi_1 \frac{\partial}{\partial \xi_1} + \dots + \xi_n \frac{\partial}{\partial \xi_n} \right)' p_{il}(\xi, \eta) \Big|_{\xi=\eta=0},$$

$$w_i^{(k)} = \{[w_{i1}, \dots, w_{im}]^T\}^{(k)} \triangleq [w'_{i1}, \dots, w'_{im}]^T,$$

$$w'_{il} \triangleq \left(\eta_1 \frac{\partial}{\partial \eta_1} + \dots + \eta_m \frac{\partial}{\partial \eta_m} \right)' p_{il}(\xi, \eta) \Big|_{\xi=\eta=0},$$

$$l = 1, \dots, n, \quad (22)$$

where subscript $\xi = \eta = 0$ means that derivatives of a term $p_{il}(\cdot, \cdot)$ are evaluated at zero values of its arguments. $P_i(\xi, \eta)$ will be referred to as an *odd r_{i1}, r_{i2} -algebraic function in the vicinity of 0* if

(1) expansion (21) may have nonzero terms only for $j \in [1, r_{i1}]$, $k \in [1, r_{i2}]$, $r_{i1}, r_{i2} < \infty$ with the last nonzero terms at $j = r_{i1}$, $k = r_{i2}$.

(2) expansion (21) has no terms with $j = 2n$, $k = 2n$, $n = 1, 2, 3, \dots$

In (21) a term $(1/j!)v_i^{(j)}$ with $\xi = y + u$ can be represented as

$$\frac{1}{j!} v_i^{(j)} \Big|_{\xi=y+u} = \frac{1}{j!} \beta_j + \frac{1}{(j-1)!} S'_j y + \text{HOT}(y) \quad (23)$$

where the elements of vector β_j are algebraic forms of order j with respect to the components of vector u , the elements of matrix S'_j are algebraic forms of order $j-1$ with respect to u , and $\text{HOT}(y)$ denotes higher order terms in y . For example, the element s'_{lm} of matrix S'_1 is

$$s'_{lm} = d'_{lm}[(d'_{l1}u_1 + \dots + d'_{ln}u_n)^2 - \sum_{r=1}^n \sum_{k=1}^n d'_{lr}u_r d'_{lk}u_k] \quad (24)$$

$$l, m = 1, \dots, n; \quad r \neq k, \quad r \neq m, \quad k \neq m$$

where

$$d'_{lm} = d'_{lm}\{p_i(\xi, \eta)\} \Big|_{\xi=\eta=0} = \frac{\partial p_{il}(\xi, \eta)}{\partial \xi_m} \Big|_{\xi=\eta=0}$$

and

$$d'_{lm}d'_{lk} = d'_{lm}d'_{lk}\{p_i(\xi, \eta)\} \Big|_{\xi=\eta=0} = \frac{\partial^2 p_{il}(\xi, \eta)}{\partial \xi_m \partial \xi_k} \Big|_{\xi=\eta=0} \quad (25)$$

Similarly, in (21) a term $(1/j!)w_i^{(j)}$ with $\eta = y + u$ can be represented as

$$\frac{1}{j!} w_i^{(j)} \Big|_{\eta=y+u} = \frac{1}{j!} \beta_j + \frac{1}{(j-1)!} E'_j y + \text{HOT}(y) \quad (26)$$

where the elements of matrix E'_j are algebraic forms of order $j-1$ with respect to u , and, for example the element e'_{lm} of matrix E'_1 is given by the right hand side of (24) and by (25) with differentiation with respect to η , i.e. the second argument of $P_i(\cdot, \cdot)$.

Let u in (23) and (26) be the zero average primitive of vector $m(t)$

$$u(t) \triangleq \int m(t) dt \quad \text{and} \quad \bar{u}(t) = 0. \quad (27)$$

Introduce a matrix

$$H_i \triangleq \frac{\partial P_i(\xi, \eta)}{\partial \xi} \Big|_{\xi=\eta=0} + \frac{\partial P_i(\xi, \eta)}{\partial \eta} \Big|_{\xi=\eta=0} + \frac{1}{2!} S'_2(u(t)) + \frac{1}{4!} S'_4(u(t)) + \dots + \frac{1}{(r_{i1}-1)!} S'_{r_{i1}}(u(t)) + \frac{1}{2!} E'_2(u(t)) + \frac{1}{4!} E'_4(u(t)) + \dots + \frac{1}{(r_{i2}-1)!} E'_{r_{i2}}(u(t)). \quad (28)$$

Theorem 2. Assume that in (6) each $P_i(\cdot, \cdot)$ is an odd r_{i1}, r_{i2} -algebraic function in a sufficiently large neighborhood of 0. Then

(i) 0 of (6) is v -stabilizable if there exists a PAZ vector $m(t)$ such that with $u(t)$ defined in

$$(27) \quad H \triangleq \sum_{i=1}^m H_i \text{ is a Hurwitz matrix;}$$

(ii) there exists positive $\epsilon_0 = \text{const.}$ such that $x_i = 0$ of (6) is v -stabilizable by vector additive vibrations $l(t) = (1/\epsilon)m(t/\epsilon)$, $0 < t \leq \epsilon_0$, where $m(t)$ satisfies all the conditions of assertion (i) above.

Corollary 2. Let an assumption of Theorem 2 hold for system (1) with fixed delays d_i , $i = 1, \dots, m$. Suppose that there exists a set of constants r_i , $i = 1, \dots, m$, such that $d_i/r_i = d_j/r_j$, $i, j = 1, \dots, m$, and $x_i = 0$ of system (6) with these constants r_i , $i = 1, \dots, m$, is v -stabilizable by vector additive vibrations $(1/\epsilon)m(t/\epsilon)$, $0 < t \leq \epsilon_0$.

Then the zero equilibrium of (1) is v -stabilizable by vector additive vibrations $(1/\epsilon_1)m(t/\epsilon_1)$, $\epsilon_1 \triangleq d_i/r_i$, if $\epsilon_1 \leq \epsilon_0$.

Remark 7. Unlike linear multiplicative vibrations vector additive vibrations are incapable of

stabilizing an unstable linear system. Indeed, in general, an unstable linear system has an unstable impulse response and is not bounded-input-bounded-state stable, therefore, any additive time-periodic input will give rise to an unbounded growth of system state. Consequently, nonlinearity is necessary for the v -stabilizability of an unstable equilibrium of a dynamical system by vector additive vibrations.

Example 2. Van der Pol equation with time delays. Consider equation

$$\begin{aligned} \ddot{x}(t) + \mu_1(x^2(t) - 1)\dot{x}(t) + b_1x(t) \\ + \mu_2(x^2(t-d) - 1)\dot{x}(t-d) \\ + b_2x(t-d) = 0 \end{aligned} \tag{29}$$

or in state space form with vector additive vibrations

$$\begin{aligned} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} &= \begin{bmatrix} x_2(t) \\ -b_1x_1(t) + \mu_1x_2(t) - \mu_1x_1^2(t)x_2(t) \\ 0 \\ -b_2x_1(t-d) + \mu_2x_2(t-d) - \mu_2x_1^2(t-d)x_2(t-d) \end{bmatrix} \\ \begin{bmatrix} l_1(t) \\ l_2(t) \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned} \tag{30}$$

Here the right hand side of (30) with $l_1(t) = l_2(t) = 0$ and d replaced by tr satisfies all assumptions of Theorem 2 and it is an odd 3,3-algebraic function around $x_{11} = x_{21} = 0$. Several trajectories of equation (30) without vibrations, i.e. with $l_1(t) = l_2(t) = 0$, shown in Fig. 3 for $b_1 = b_2 = \mu_1 = \mu_2 = 0.5$ demonstrate an instability of an equilibrium $x_i = 0$. Choosing $m_1(t) = \alpha \cos t$ and $m_2(t) = 0$, matrix H of Theorem 2 is given by

$$H = \begin{bmatrix} 0 & 1 \\ -(b_1 + b_2) & -(\mu_1 + \mu_2)\left(1 - \frac{\alpha^2}{2}\right) \end{bmatrix} \tag{31}$$

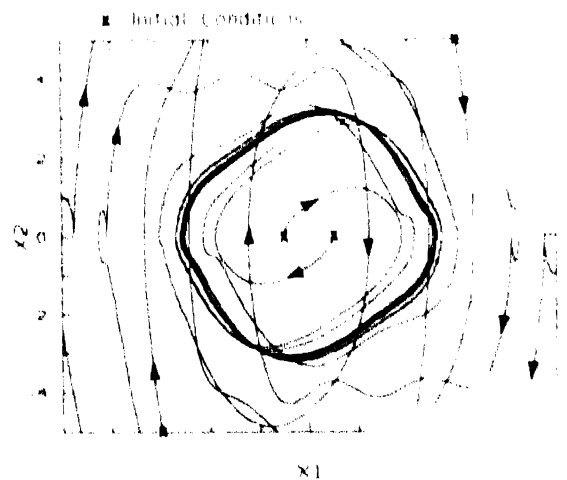


FIG. 3. Phase portrait of Van der Pol equation with a time lag $d = 0.4\pi$.

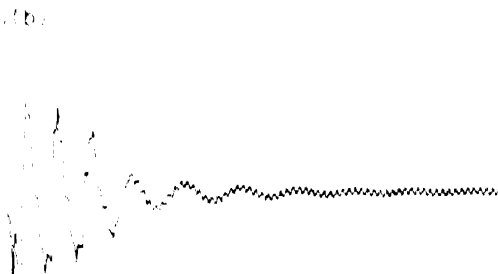
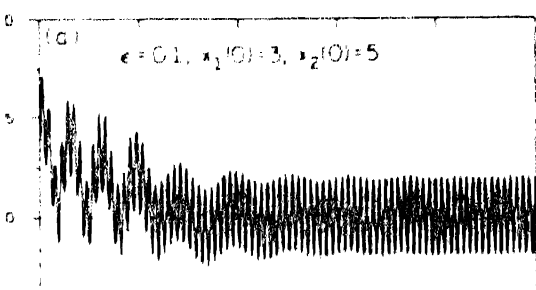


FIG. 4. Solutions $x_1(t)$ and $x_2(t)$ versus time of Van der Pol equation with oscillations and a time lag $d = 0.4\pi$.

Hence for $(b_1 + b_2) > 0$ and $(\mu_1 + \mu_2) > 0$ zero equilibrium of the Van der Pol equation with delays (29) is v -stabilizable by vector additive vibrations $l_1(t) = (\alpha/\epsilon) \cos(t/\epsilon)$, $l_2(t) = 0$, if $\alpha > \sqrt{2}$ and delay d is sufficiently small. Such stabilization is indeed shown in Fig. 4 for $d = 0.4\pi$. The last element in the lower row of matrix H in (31) demonstrates that nonlinearities play a decisive role in the vibrational stabilization of this class of systems by vector additive vibrations.

III. TRANSIENT BEHAVIOR ANALYSIS OF VIBRATIONALLY CONTROLLED NONLINEAR TIME LAG SYSTEMS

Vibrational stabilization considered in the previous sections addresses the local behavior of vibrationally controlled nonlinear systems around an equilibrium point with the main emphasis on attractivity, i.e. on behavior as $t \rightarrow \infty$. For control purposes it is also of interest to analyze the nonlocal system behavior at every time moment starting from $t = 0$, i.e. the transient behavior of a system. Analysis of trajectories of vibrationally controlled system (3) is a difficult

task; however, if vibrations are of the form $f(t) = (1/\varepsilon)\phi(t/\varepsilon)$, $\phi(\cdot)$ -PAZ function, $0 < \varepsilon \ll 1$, $\varepsilon = \text{const.}$, i.e. sufficiently fast, then trajectories of system (3) are usually composed of a fast oscillatory part with the period of $\phi(t/\varepsilon)$ superimposed on a slow part. A comparison of a slow part of a trajectory of the oscillatory system with a trajectory of the corresponding system without vibrations for the same initial conditions reveals the qualitative changes in system behavior caused by oscillations.

In order to analyze the transient behavior of a vibrationally controlled system, a solution of the initial value problem for every delay equation in this paper will be denoted as $x(t, x_0, 0)$ and interpreted in the sense of (Driver, 1977, p. 257) as a continuous function $x: [-r, \infty) \rightarrow R^n$ that reproduces the initial data, curve $x_0(s)$, $s \in [-r, 0]$, and satisfies the equation considered for $t \geq 0$ with $\dot{x}(0)$ being understood as the right-hand derivative.

In this section we consider the transient behavior analysis of the system (6) with linear multiplicative and vector additive vibrations on a finite time interval. Then we relate it to system (4) with fixed delays, d_i , $i = 1, \dots, m$, as in the previous section. Thus we consider system

$$\dot{x}(t) = \sum_{i=1}^m P_i(x(t), x(t - \tau_i), \lambda_0) + Q((1/\varepsilon)\phi(t/\varepsilon), x(t)) \quad (32)$$

with the right-hand side defined in (1) and (4) and τ_i , $i = 1, \dots, m$, being positive constants of the order $O(1)$.

Let $x(t, x_0, 0)$, where $x(0, x_0, 0) = x_0$ be a trajectory of equation (32). In order to strip $x(t, x_0, 0)$ of its fast oscillating component introduce a *moving average along a trajectory* $x(t, x_0, 0)$ as

$$\bar{x}(t) \triangleq \frac{1}{T} \int_t^{t+T} x(s, x_0, 0) ds, \quad 0 \leq t < \infty, \quad (33)$$

where T is a period of $\phi(t/\varepsilon)$.

If the quantity $\bar{x}(t)$ can be closely approximated by the trajectory of a time-invariant system then the transient behavior analysis of system (32) for various magnitudes and frequencies of the oscillations can be greatly simplified, resulting in the constructive procedure for the design of the parametric excitations that induce the desired qualitative changes in the system behavior.

A. Linear multiplicative vibrations

Let linear multiplicative vibrations in (32) be of the form $(1/\varepsilon)F(t/\varepsilon)x(t)$. Assume that

equation

$$\dot{x}(t) = F(t)x(t) \quad (34)$$

has a periodic in t fundamental matrix $\Phi(t)$.

Along with (33) introduce

$$\bar{x}_M(z(t)) \triangleq \Phi(t)z(t, z_0, 0), \quad (35)$$

where $\Phi(t) = (1/T^*) \int_0^t \Phi(\tau) d\tau$, $T^* \triangleq T/\varepsilon$, and $z(t, z_0, 0)$ with $z(0, z_0, 0) = z_0 = \text{const.}$ is a solution of the equation

$$\dot{z}(t) = \bar{Q}(z(t)), \quad \bar{Q}(y) \triangleq \lim_{T \rightarrow \infty} (1/T) \int_0^T Q(t, y) dt \quad (36)$$

where

$$Q(t, y) \triangleq \Phi^{-1}(t) \sum_{i=1}^m P_i(\Phi(t)y, \Phi(t - \tau_i)y, \lambda_0). \quad (37)$$

If $\bar{x}_M(z(t))$ stays close to $\bar{x}(t)$ on a time interval of interest, then $\bar{x}_M(z(t))$ can be viewed as an *approximate moving average along a trajectory* $x(t, x_0, 0)$ of (32) with linear multiplicative vibrations on this time interval. Theorem 3 below gives the conditions under which $\bar{x}_M(z(t))$ and $\bar{x}(t)$ can be made arbitrarily close on the arbitrarily large finite time interval.

Theorem 3. Assume that (a) functions $P_i(\zeta, \eta, \lambda_0)$, $i = 1, \dots, m$, of system (32) are continuously differentiable with respect to $\zeta, \eta \in \Omega_1 \subset R^n$ and (b) fundamental matrix $\Phi(t)$ of (34) is T^* -periodic where T^* is a period of $F(\cdot)$ in (34).

Then for any δ as small as desired and κ as large as desired there exists $\varepsilon_0 = \varepsilon_0(\delta, \kappa)$ such that for $0 < \varepsilon \leq \varepsilon_0$ the following holds

$$\|\bar{x}(t) - \bar{x}_M(z(t))\| < \delta, \quad t \in [0, \kappa], \quad (38)$$

where $z_0 = \Phi(0)x_0(0)$.

Corollary 3. Let assumptions (a) and (b) of Theorem 3 hold for system (4) with fixed delays d_i , $i = 1, \dots, m$, and linear multiplicative vibrations $D(t)x(t)$. Suppose that there exists a set of constants r_i , $i = 1, \dots, m$, such that $d_i/r_i = d_j/r_j$, $i, j = 1, \dots, m$, and solutions of system (32) with these constants r_i satisfy inequality (38) for given δ and κ for $0 < \varepsilon \leq \varepsilon_0$. Then solutions of system (4) with vibrations $D(t)x(t) = (1/\varepsilon_1)F(t/\varepsilon_1)x(t)$, $\varepsilon_1 \triangleq d_i/r_i$, satisfy inequality (38) for the same δ and κ if $\varepsilon_1 \leq \varepsilon_0$.

Example 3. Duffing equation (15) with time lags and linear multiplicative vibrations

$$\frac{1}{\varepsilon} F\left(\frac{t}{\varepsilon}\right)x = \begin{bmatrix} 0 & 0 \\ \frac{\alpha}{\varepsilon} \cos \frac{t}{\varepsilon} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (39)$$

Since $\Phi(t)$ in this case is given by (19), from (35) we have

$\bar{x}_{M1}(z(t)) = z_1(t) \text{ and } \bar{x}_{M2}(z(t)) = z_2(t) \tag{40}$

where $z(t) \triangleq z(t, z_0, 0)$ is a solution of equation (36) which for this specific case takes the form.

$$\begin{aligned} \dot{z}_1(t) &= z_2(t), \\ \dot{z}_2(t) &= -(a_1 + a_2)z_2(t) \\ &\quad + \left(b_1 + b_2 - \frac{\alpha^2}{2}\right)z_1(t) - (c_1 + c_2)z_1^3(t). \end{aligned} \tag{41}$$

Figure 5 shows that $\bar{x}_{M1}(z(t))$ and $\bar{x}_{M2}(z(t))$ given by dashed curves indeed represent approximate moving averages along $x_1(t)$ and $x_2(t)$, respectively, given by solid curves for delay $d = 0.1\pi$, $\epsilon = 0.05$, $a_1 = 0.6$, $a_2 = 0.4$, $b_1 = b_2 = 0.5$, and $c_1 = c_2 = 0.05$.

B. Vector additive vibrations

Let vector additive vibrations in (32) be of the form $(1/\epsilon)m(t/\epsilon)$.

Theorem 4. Let assumption (a) of Theorem 3 hold and $u(t)$ be the T^* -periodic zero mean primitive of $m(t)$.

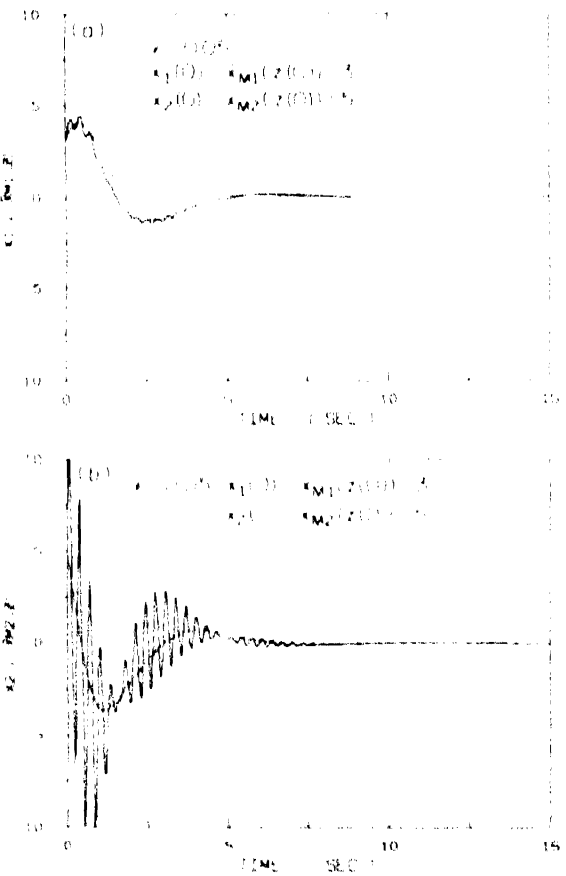


FIG. 5. Solutions $x_1(t)$ and $x_2(t)$ versus time of Duffing equation with oscillations and a time lag $d = 0.1\pi$ and their approximate moving averages $\bar{x}_{M1}(z(t))$ and $\bar{x}_{M2}(z(t))$ versus time

Then for any δ as small as desired and κ as large as desired there exists $\epsilon_0 = \epsilon_0(\delta, \kappa)$ such that for $0 < \epsilon \leq \epsilon_0$ the following holds

$\|\bar{x}(t) - z(t)\| < \delta, \quad t \in [0, \kappa] \tag{42}$

where $z_0 = u(0) + x_0(0)$, and $z(t, z_0, 0)$ is a solution of an ordinary differential equation

$$\begin{aligned} \dot{z}(t) &= \frac{1}{T^*} \int_0^{T^*} \sum_{i=1}^m P_i[z(t) + u(s), z(t) \\ &\quad + u(s - \tau_i), \lambda_0] ds. \end{aligned} \tag{43}$$

Corollary 4. Let all assumptions of Theorem 4 hold for system (4) with fixed delays d_i , $i = 1, \dots, m$, and vector additive vibrations $l(t)$. Suppose that there exists a set of constants r_i , $i = 1, \dots, m$, such that $d_i/r_i = d_j/r_j$, $i, j = 1, \dots, m$ and solutions of system (32) with these constants r_i satisfy inequality (42) for given δ and κ for $0 < \epsilon \leq \epsilon_0$. Then solutions of system (4) with vibrations $l(t) = (1/\epsilon_1)m(t/\epsilon_1)$, $\epsilon_1 \triangleq d_i/r_i$, satisfy inequality (42) for the same δ and κ if $\epsilon_1 \leq \epsilon_0$.

Example 4. Van der Pol equation (30) with time lags and vector additive vibrations

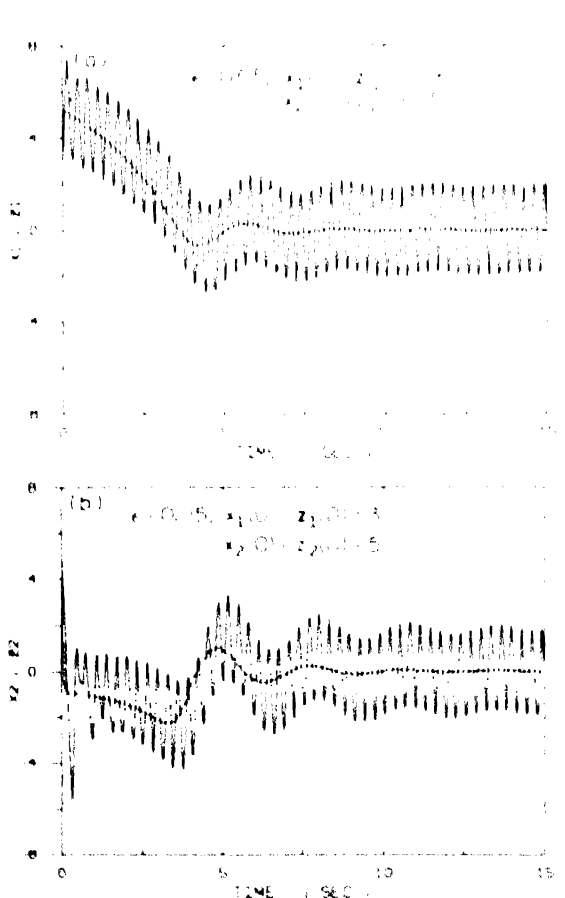


FIG. 6. Solutions $x_1(t)$ and $x_2(t)$ versus time of Van der Pol equation with oscillations and a time lag $d = 0.2$ and their approximate moving average, $\bar{x}_1(t)$ and $\bar{x}_2(t)$ versus time.

$[(\alpha/\epsilon) \sin(t/\epsilon), (\alpha/\epsilon) \sin(t/\epsilon)]^T$. Equation (43) in this case has the form

$$\begin{aligned}\dot{z}_1(t) &= z_2(t), \\ \dot{z}_2(t) &= [(b_1 + b_2) + (\mu_1 + \mu_2)\alpha^2]z_1(t) \\ &\quad + (\mu_1 + \mu_2)\left(1 - \frac{\alpha^2}{2}\right)z_2(t) \\ &\quad - (\mu_1 + \mu_2)z_1^2(t)z_2(t).\end{aligned}\quad (44)$$

Approximate moving averages $z_1(t)$ and $z_2(t)$ (dashed curves) along with the corresponding solutions $x_1(t)$ and $x_2(t)$ (solid curves), respectively, are shown in Fig. 6 for $b_1 = b_2 = \mu_1 = \mu_2 = 0.5$, $d = 0.2$ and $\epsilon = 0.05$.

IV. CONCLUSIONS

This paper demonstrates that under certain conditions, parametric vibrations approximately introduced into a nonlinear time lag system of the form (1) are capable of converting an unstable equilibrium of a system into an asymptotically stable one or creating an asymptotically stable oscillatory regime with the average located at an unstable equilibrium point. The criteria presented enable one to investigate the existence of the stabilizing vibrations and give procedures for the choice of their parameters. While the theory presented is restricted to the delays of the order $O(\epsilon)$, $0 < \epsilon \ll 1$, Examples 1 and 2 demonstrate that vibrations are capable of stabilizing systems with delays of the order $O(1)$. The method for the transient behavior analysis of a vibrationally controlled system is also given and is supported by numerical examples. Thus, vibrational control is shown to be a possible alternative for control of nonlinear systems with time delays in the situations where conventional methods are expensive, difficult, or impossible to apply due to restrictions on sensing and actuation.

Acknowledgements—This work was supported in part by the National Science Foundation under the Presidential Young Investigator Award Grant MSS-8957198 and in part by the National Center for Supercomputing Applications, grant no. ECS890023N and utilized the CRAY X-MP/48 system at the National Center for Supercomputing Application at the University of Illinois at Urbana-Champaign.

REFERENCES

- Bellman, R., J. Bentsman and S. M. Meerkov (1985). Stability of fast periodic systems. *IEEE Trans. Aut. Control*, **AC-30**, 289–291.
- Bellman, R., J. Bentsman and S. M. Meerkov (1986a). Vibrational control of nonlinear systems. *Vibrational stabilizability*. *IEEE Trans. Aut. Control*, **AC-31**, 710–716.
- Bellman, R., J. Bentsman and S. M. Meerkov (1986b). Vibrational control of nonlinear systems. *Vibrational controllability and transient behavior*. *IEEE Trans. Aut. Control*, **AC-31**, 717–724.
- Bentsman, J. (1987). Vibrational control of a class of nonlinear systems by nonlinear multiplicative vibrations. *IEEE Trans. Aut. Control*, **AC-32**, 711–716.
- Bentsman, J. and H. Hvoslov (1988). Vibrational control of a laser illuminated thermochemical system. *ASME J. Dynam. Syst. Meas. Control*, **110**, 109–113.
- Brockett, R. W. (1970). *Finite Dimensional Linear Systems*. Wiley, New York.
- Cinar, A., J. Deng, S. M. Meerkov and X. Shu (1987). Vibrational control of an exothermic reaction in a CSTR. Theory and experiment. *AIChE J.*, **33**, 353–365.
- Driver, R. D. (1977). *Ordinary and Delay Differential Equations*. Springer, New York.
- Fakhtakh, J. and J. Bentsman (1990). Experiment with vibrational control of a laser illuminated thermochemical system. *ASME J. Dyn. Syst., Meas. Control*, **112**, 42–47.
- Halanay, A. (1966). *Differential Equations: Stability, Oscillations, Time Lags*. Academic Press, New York.
- Hale, J. (1966). Averaging methods for differential equations with retarded arguments and a small parameter. *J. Diff. Eqs.*, **2**, 57–73.
- Khargonekar, P., K. Poolla and A. Tannenbaum (1985). Robust control of linear time-invariant plants using periodic compensation. *IEEE Trans. Aut. Control*, **AC-30**, 1088–1096.
- Kolmanovskii, V. B. and V. R. Nosov (1986). *Stability of Functional Differential Equations*. Academic Press, Florida.
- Lee, S., S. M. Meerkov and T. Runolfsson (1987). Vibrational-feedback control: Zero placement capabilities. *IEEE Trans. Aut. Control*, **AC-32**, 604–611.
- Ray, W. H. (1981). *Advanced Process Control*. McGraw-Hill, New York.
- Strizak, I. G. (1982). *Method of Averaging in the Problems of Mechanics*. Vishcha Shkola, Kiev-Donetsk, USSR.

APPENDIX. PROOFS OF THE FORMAL STATEMENTS

Proof of Theorem 1. With vibrations

$$Q(f(t), x(t)) \sim D(t)x(t) \sim (1/\epsilon)F(t/\epsilon)x(t), \quad (1a)$$

$\tau \triangleq t/\epsilon$, and $d_i \rightarrow \tau_i$, $i = 1, \dots, m$, equation (4) takes the form

$$\begin{aligned}\frac{dx(\tau)}{d\tau} &= \epsilon \sum_{i=1}^m P_i(x(\tau), x(\tau - \tau_i), \lambda_0) \\ &\quad + F(\tau)x(\tau)\end{aligned}\quad (2a)$$

Introducing into equation (2a) coordinate transformation

$$x(\tau) = \Phi(\tau)y(\tau), \quad (3a)$$

which for the delayed states takes the form

$$x(\tau - \tau_i) = \Phi(\tau - \tau_i)y(\tau - \tau_i), \quad i = 1, \dots, m, \quad (4a)$$

(2a) reduces to a standard form

$$\begin{aligned}\frac{dy(\tau)}{d\tau} &= \epsilon \Phi^{-1}(\tau) \sum_{i=1}^m \\ &\quad \times P_i(\Phi(\tau)y(\tau), \Phi(\tau - \tau_i)y(\tau - \tau_i), \lambda_0)\end{aligned}\quad (5a)$$

Since $F(\tau)$ is a periodic bounded zero mean function of τ , by Abel's formula (cf. Brockett, 1970, Theorem 3.3) $\Phi^{-1}(\tau)$ is bounded for $\tau \in (-\infty, \infty)$ and the right-hand side of (5a) is well defined for all τ .

Averaging the right hand side of (5a) with respect to τ , linearizing it at $y(\tau) = y(\tau - \tau_i) \approx 0$ and dropping the delays in the states we obtain an ordinary differential equation

$$\frac{dy(\tau)}{d\tau} = \epsilon \left[\Phi^{-1}(\tau)A\Phi(\tau) + \sum_{i=1}^m \Phi^{-1}(\tau)B_i\Phi(\tau - \tau_i) \right] y(\tau) \quad (6a)$$

that for sufficiently small ϵ governs the stability properties of the trivial solution of (5a). Finally, noting that the averaged equation corresponding to (5a) is given in time t by $\dot{z}(t) = Rz(t)$, where matrix R is defined in (10), and that if $\Phi(\tau)$ is almost periodic then (3a) and (4a) are stability

preserving, the assertions of the theorem directly follow from Theorem 3.3 of Hale (1966) Q.E.D.

Proof of Corollary 1. Under the assumptions of Corollary 1, conditions of assertion (i) of Theorem 1 are satisfied for systems (1). Therefore, for $\epsilon_1 < \epsilon_0$, the proof of Corollary 1 directly follows from assertion (ii) of Theorem 1 Q.E.D.

Proof of Theorem 2. System (6) with vector additive vibrations $l(t) = (1/\epsilon)m(t/\epsilon)$ in time $\tau \leq t/\epsilon$ is given by

$$\frac{dx(t)}{dt} = \epsilon \sum_{i=1}^m P_i(x(\tau), x(\tau - \tau_i), \lambda_0) + m(\tau) \tag{7a}$$

With $u(\tau)$ defined by (27), substitutions

$$x(\tau) = y(\tau) + u(\tau) \tag{8a}$$

and

$$x(\tau - \tau_i) = y(\tau - \tau_i) + u(\tau - \tau_i) \tag{9a}$$

reduce (7a) to a standard form

$$\begin{aligned} \frac{dy(\tau)}{d\tau} = & \epsilon \sum_{i=1}^m P_i[y(\tau) + u(\tau), y(\tau - \tau_i) \\ & + u(\tau - \tau_i), \lambda_0] \end{aligned} \tag{10a}$$

Expanding every $P_i(\xi, \eta, \lambda_0) = P_i(\xi, \eta)$ around $\xi = \eta = 0$, under the assumption of Theorem 2, we obtain

$$\begin{aligned} \frac{dy(\tau)}{d\tau} = & \epsilon \sum_{i=1}^m \left\{ \sum_{j=1}^{2k} \frac{1}{j!} v_j^{(i)} |_{\xi=y(\tau), \eta=y(\tau)} u(\tau) \right. \\ & \left. + \sum_{j=1}^{2k} \frac{1}{j!} w_j^{(i)} |_{\eta=y(\tau - \tau_i), \xi=u(\tau - \tau_i)} \right\} \end{aligned} \tag{11a}$$

Averaging the right hand side of (11a) with respect to τ and dropping the delays in the states, we have

$$\frac{dz(\tau)}{d\tau} = \left(\sum_{i=1}^m H_i \right) z(\tau) + \text{HOT}(z) \tag{12a}$$

where matrices $H_i, i = 1, \dots, m$, are defined in (28)

Assume now that $\sum_{i=1}^m H_i$ is a Hurwitz matrix. Then, noting that (12a) has a zero equilibrium, by Theorem 3.3 of Hale (1966) for every $\delta > 0$ there exists $\epsilon_0(\delta)$ such that (10a) with $0 < \epsilon \leq \epsilon_0$ has a unique asymptotically stable almost periodic solution $y^1(\tau)$ characterized by

$$\|y^1(\tau)\| \leq \delta \tag{13a}$$

Now, the proof Theorem 2 follows from (13a) upon noting that due to (8a)

$$x^1(\tau) = y^1(\tau) + u(\tau) = \bar{y}^1(\tau)$$

Q.E.D.

Proof of Corollary 2 Directly follows from Theorem 2

Proof of Theorem 3 Follows that of Theorem 1 until and

including the sentence after equation (5a). Dropping the delays in state variable $y(\cdot)$ and averaging the right hand side of (5a) we obtain

$$\frac{dz(\tau)}{d\tau} = \epsilon \bar{Q}(z(t)) \tag{14a}$$

where $\bar{Q}(\cdot)$ is given in (36) and (37).

Now by Theorem 4.32 of Halanay (1966, p. 460) for $y(0) = y_0(0) = z(0) = z_0$ and every $\kappa > 0$ and $\delta > 0$ there exists $\epsilon_2 > 0$ such that for $0 < \epsilon \leq \epsilon_2$ we have

$$\|y(\tau, y_0, 0) - z(\tau, z_0, 0)\| \leq \eta, \quad \forall \tau \in [0, \kappa/\epsilon] \tag{15a}$$

Since

$$x(t, x_0, 0) = \Phi(t/\epsilon)y(t, x_0, 0), \tag{16a}$$

$T = O(\epsilon)$, and $\bar{x}(t)$ and $\bar{y}(t)$ are of the order $O(1)$ in time $t \approx \tau/\epsilon$, we have

$$\begin{aligned} \bar{x}(t) &= \frac{1}{T} \int_0^T \Phi(t/\epsilon)y(t) dt \\ &= \left[\frac{1}{T} \int_0^T \Phi\left(\frac{t}{\epsilon}\right) dt \right] y(t) + K_1(t) \\ &= \left[\frac{1}{T} \int_0^T \Phi(s) ds \right] y(t) + K_1(t) \\ &= \bar{\Phi}(t)y(t) + K_1(t), \end{aligned} \tag{17a}$$

where $y(t) = y(t, x_0, 0)$, $K_1(t) = O(\epsilon)$ and $\|K_1(t)\|$ uniformly approaches 0 as $\epsilon \rightarrow 0$.

Denoting $z(t) = z(t, z_0, 0)$, for the left hand side of inequality (38) from (15a) and (17a) for any given η and κ we have in time t

$$\begin{aligned} \|\bar{x}(t) - \bar{x}_M(z(t))\| &= \|\bar{\Phi}(t)y(t) + K_1(t) - \bar{\Phi}(t)z(t)\| \\ &\leq \|\bar{\Phi}(t)\| \|y(t) - z(t)\| + \|K_1(t)\| \\ &\leq N \|y(t) - z(t)\| + \|K_1(t)\| \\ &\leq N\eta + \|K_1(t)\|, \quad 0 < t \leq \epsilon_j, \end{aligned} \tag{18a}$$

where N is some positive constant which exists since $\bar{\Phi}(t)$ is a constant matrix with bounded elements

Finally, since we can always choose $\epsilon_0 < \epsilon_2$ and η such that

$$N\eta + \|K_1(\epsilon_0)\| \leq \delta, \quad \forall t \in [0, \kappa],$$

Theorem 3 is proven

Q.E.D.

Proof of Corollary 3. Directly follows from Theorem 3.

Proof of Theorem 4. Follows that of Theorem 2 until and including equation (10a). Dropping the delays in state variable $y(\cdot)$ and averaging the right hand side of (10a) we obtain in time t equation (43). Now the proof follows that of Theorem 3 after equation (14a), with $\Phi(t/\epsilon)y(t, x_0, 0)$ replaced by $u(t/\epsilon) + y(t, x_0, 0)$ and $\bar{x}_M(z(t))$ replaced by $z(t)$. Q.E.D.

Proof of Corollary 4. Directly follows from Theorem 4.

Generalized Zero Sets Location and Absolute Robust Stabilization of Continuous Nonlinear Control Systems*

GILA FRUCHTER†

An analytic method for finding the location of the generalized zero set of a vector-valued function is applied here to find the complete feasible set of sectors of nonlinearities which allows absolute robust stability of a Lurie type system with linear part under uncertainty conditions, according to the Popov criterion.

Key Words—Control system design, nonlinear control systems, robust control, stability, zeros

Abstract—We describe an analytic method for finding the location of the generalized zero set of a vector-valued function which depends on m real variables and $(n+k)$ complex parameters. The method is applied to a robust design problem of a nonlinear Lurie type continuous-time system, with the linear part under uncertainty conditions. We find the complete feasible set of sectors of the nonlinearities which allows robust absolute stability of the system according to the Popov criterion. Illustrative numerical examples are provided.

1. INTRODUCTION

VARIOUS PROBLEMS which arise in system theory can be reduced to the problem of locating the generalized zero set, defined as follows.

A *generalized zero set*. Given an open set G in \mathbb{R}^m , $m \geq 1$, closed sets Q_1, \dots, Q_n , P_1, \dots, P_k in $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ with $Q = Q_1 \times \dots \times Q_n$, $P = P_1 \times \dots \times P_k$ and a continuous function $f: Q \times P \times G \rightarrow \mathbb{R}^d$, the generalized zero set of f relative to Q , P and G is defined by

$$M = \{s \in G \subset \mathbb{R}^m : f(q, p, s) = 0 \text{ for some } q \text{ in } Q \text{ and all } p \text{ in } P\}.$$

If we let $f_{q,p}(s) = f(q, p, s)$, $q \in Q$, $p \in P$, $s \in G$, then

$$M = \bigcap_{p \in P} \bigcup_{q \in Q} f_{q,p}^{-1}(0).$$

In other words the set M can be written as

*Received 23 January 1990, revised 6 July 1990, received in final form 17 September 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor P. Dorato under the direction of Editor H. Kwakernaak.

†Department of Mathematics, Technion-Israel Institute of Technology, Haifa, 32000 Israel. Currently with the Dept. of Aerospace, University of California Los Angeles, Los Angeles, CA 90024, U.S.A.

intersections of sets

$$V_p = \bigcup_{q \in Q} f_{q,p}^{-1}(0), \quad p \in P,$$

where V_p are zero sets (Fruchter, 1988; Fruchter *et al.*, 1987a, 1991a, b) recently developed into a tool and used in solving problems in control theory. Establishing an algorithm to finding M makes more applications possible in a wide variety of areas.

Walach and Zehavi (1982) consider this problem for $m=2$ and $G=\mathbb{R}^2$, and when $f(q, p, s)$ is a complex-valued function, holomorphic in q and p and a polynomial in the complex variable $s=s_1+j s_2$, and when each Q_i and P_j is either a piecewise-smooth simple arc or a Jordan domain bounded by a piecewise-smooth curve. They were able to find necessary conditions for the boundary of M and thereby to develop an algorithm for determining M in some cases. Their results have already been instrumental in control theory.

Here, we derive a new algorithm for locating M for continuously differentiable vector-valued functions f and quite general parameter spaces Q and P .

The method is applied to a problem of robustness of absolute stability of Lurie type (Lurie, 1954) continuous nonlinear systems, with linear parts under uncertainty conditions. By uncertainty we mean that the transfer function of the linear part may have as coefficients parameters which may take values in a given set. We find *all* the feasible set of sectors of nonlinearities, for which absolute robust stability of the system under uncertainty conditions in the

linear part is ensured, according to the Popov criterion (Popov, 1961). From the engineering standpoint, such a result will provide the designer of such a system with an excellent design flexibility, in choosing the *optimal* robust sector considering design constraints.

A partial solution to this problem was given in Siljak's work (Siljak, 1969a, b, 1989). Siljak's method gives sufficient conditions for absolute robust stability, according to Popov criterion, for a restricted class of transfer functions, with respect to the form of the appearance of the uncertainty parameters. This is the price of using Kharitonov type solution by imbedding polytopes of dimensions equal to the number of the uncertainty parameters plus two, in the solution set.

The method proposed in this paper gives *sufficient and necessary* conditions for absolute robust stability with respect to the uncertainties in the linear part, according to Popov criterion. The method is *unrestricted* to the form in which the uncertainty parameters may appear in the transfer function. The complete set of sectors of the nonlinearities is selected from the *two-dimensional* set:

$$\mathcal{P} = \mathbb{R}^* \times \mathbb{R} \setminus V = \{(k, q) \in \mathbb{R}^* \times \mathbb{R} : k^{-1} + \operatorname{Re}[(1 + j\omega q)G(j\omega, r)] > 0, \forall \omega \geq 0, \forall r \in \mathcal{P}/\mathcal{V}\},$$

where r is the vector of all uncertainty parameters and \mathcal{P}/\mathcal{V} denotes the parameter space in which r may take its values. The set V is a corresponding two-dimensional zero set of points (k, q) in $\mathbb{R}^* \times \mathbb{R}$. It is determined by the analytic methods described in (Fruchter *et al.*, 1987a, 1988, 1991a, b) and illustrated in this paper. The two-dimensional geometric interpretation can be done at *any number of uncertainty parameters*.

The discrete case of this problem was treated in the previous works (Fruchter *et al.*, 1987b, Fruchter, 1988).

This paper consists of five sections. In Section 2 we present a brief summary on the method of "zero sets location" (Fruchter, 1988; Fruchter *et al.*, 1987a, 1991a, b). In Section 3 we describe a method of finding the generalized zero set M . The method is illustrated by two numerical examples with engineering meaning. In Section 4 we utilize the algorithm in absolute robust stabilization of continuous-time Lurie type nonlinear systems. A numerical example is provided. The paper closes in Section 5 with appropriate concluding remarks.

2. PRELIMINARY RESULTS

We present a brief summary of the method "zero sets location" (Fruchter, 1988; Fruchter *et*

al., 1987a, 1991a, b), including notations and results needed in this paper.

A. The zero set

Definition 1 (Fruchter *et al.*, 1991a). Let $K = K_1 \times \cdots \times K_n$ be a set in $\mathbb{C}^n = \mathbb{C} \times \cdots \times \mathbb{C}$ (n times), where each K_i is a closed set in $\mathbb{C} = \mathbb{C} \cup \{\infty\}$ whose boundary ∂K_i is a finite union of piecewise-smooth simple curves and piecewise-smooth closed simple curves. Let G be an open set in \mathbb{R}^m and let $f: K \times G \rightarrow \mathbb{R}^d \cup \{\infty\}$ be a continuously differentiable function. The zero set of $f = (f_1, \dots, f_d)$ relative to K and G is then defined by

$$V = \{s \in G \subset \mathbb{R}^m : \exists A \in K \text{ such that } f(A, s) = 0\}.$$

Lemma (Fruchter *et al.*, 1991a). V is a closed set relative to G .

B. Zero sets location—Outline of the procedure

There are two principal stages in locating V . In the first stage one finds an $(m-1)$ -dimensional set L , $L \subset V$, which contains the boundary of V relative to G . Next, by picking a point in each of the connected components of $G \setminus L$ one checks which of these connected components are included in V and which are outside V . This stage leads to the original problem in a lower dimension, which can be carried out reductively until the final solution is reached. Let D_1, \dots, D_k be the connected components of $G \setminus L$, which are included in V ; then

$$V = \bigcup_{i=1}^k D_i \cup L.$$

In order to find the set L one needs necessary conditions on the boundary of V . In the following we present two theorems which provide such necessary conditions.

C. Necessary conditions on the relative boundary of the zero set

Theorem 1 (Fruchter *et al.*, 1991a). Let K , G , f and V be as in Definition 1. Suppose that s^0 is a point in the boundary of V relative to G and $A^0 = (A_1^0, \dots, A_n^0)$ is a point in K such that $f(A^0, s^0) = 0$. Suppose that for $i = 1, \dots, \ell$, $1 \leq \ell \leq n$, A_i^0 belongs to the boundary of a connected component of K_i having a parametric representation

$$A_i = A_i(\theta_i), \quad 0 \leq \theta_i \leq 1$$

and that for $i = \ell + 1, \dots, n$, $0 \leq \ell \leq n$, $A_i^0 \in \operatorname{int} K_i$.

Suppose that $\theta_1^0, \dots, \theta_\ell^0$ are numbers in the open intervals $(0, 1)$ such that

$$A_i(\theta_i^0) = A_i^0$$

and x_i^0, y_i^0 are real numbers such that for $i = \ell + 1, \dots, n$, $A_i^0 = x_i^0 + jy_i^0$.

Let $\delta \in (0, 1)$ be such that $(\theta_i^0 - \delta, \theta_i^0 + \delta) \subset (0, 1)$ for $i = 1, \dots, \ell$ and $(x_i^0 - \delta, x_i^0 + \delta) \times (y_i^0 - \delta, y_i^0 + \delta) \subset K_i$ for $i = \ell + 1, \dots, n$. Denote

$$\begin{aligned}\xi^0 &= (\xi_1^0, \dots, \xi_r^0) \\ &= (\theta_1^0, \dots, \theta_\ell^0, x_{\ell+1}^0, y_{\ell+1}^0, \dots, x_n^0, y_n^0)\end{aligned}$$

and

$$\begin{aligned}\xi &= (\xi_1, \dots, \xi_r) \\ &= (\theta_1, \dots, \theta_\ell, x_{\ell+1}, y_{\ell+1}, \dots, x_n, y_n)\end{aligned}$$

for each point $\lambda \in \mathbb{R}^r$ such that

$$|\xi_i - \xi_i^0| < \delta, \quad i = 1, \dots, r.$$

Then, at points (A^0, s^0) , where the derivatives of $A_i(\theta_i)$ exist, we have the relations

$$(i) \quad f_i(A^0, s^0) = 0, \quad i = 1, \dots, d$$

$$(ii) \quad \frac{\partial(f_1, \dots, f_d)}{\partial(\xi_{i_1}, \dots, \xi_{i_d})}(A^0, s^0) = 0, \\ 1 \leq i_1 < \dots < i_d$$

At points where any of the coordinates is ∞ , the differentiability of f and the conditions (i) and (ii) should be evaluated after suitable changes of variables of the form $z \rightarrow 1/z$ are performed.

Theorem 2 (Fruchter *et al.*, 1991a). Let $K, G, f = (f_1, \dots, f_d), V, (A^0, s^0), \xi^0 = (\xi_1^0, \dots, \xi_r^0)$ and $\xi = (\xi_1, \dots, \xi_r), r \geq 2$, with $A = A(\xi)$ and $A^0 = A(\xi^0)$ be as in Theorem 1, when $d \geq 2$.

Let k and q be integers such that $1 \leq k \leq d-1$ and $d-k \leq q \leq r-1$. Suppose that there is a $(m+r-q)$ -dimensional neighborhood $N = N_1 \times N_2$ of the point $(\xi^{1,0}, s^0) = (\xi_{q+1}^0, \dots, \xi_r^0, s^0)$ and a vector-valued function $q: N \rightarrow \mathbb{R}^q$, such that

$$(i) \quad q \text{ is continuously differentiable on } N$$

$$(ii) \quad q(\xi^{1,0}, s^0) = (\xi_1^0, \dots, \xi_q^0)$$

$$(iii) \quad v(A(q(\xi^1, s), \xi^1), s) = 0 \text{ for every } (\xi^1, s) \text{ in } N,$$

where

$$\xi^1 = (\xi_{q+1}, \dots, \xi_r)$$

and

$$v = (f_{k+1}, \dots, f_d).$$

Let $g: N \rightarrow \mathbb{R}^k$ be defined by

$$g(\xi^1, s) = u(A(q(\xi^1, s)\xi^1), s)$$

where $u = (f_1, \dots, f_k)$ and $g = (g_1, \dots, g_k)$.

Then at points $(\xi^{1,0}, s^0)$ we have

$$g_j(\xi^{1,0}, s^0) = 0, \quad j = 1, \dots, k$$

$$\frac{\partial(g_1, \dots, g_k)}{\partial(\xi_{i_1}, \dots, \xi_{i_k})}(\xi^{1,0}, s^0) = 0,$$

$$q+1 \leq i_1 < \dots < i_k \leq r.$$

D. An outline of the procedure of determining L

Let f, K and G be as in Theorem 1 and Theorem 2. For each ℓ in $\{1, \dots, n\}$, choose a subset of ℓ different indices from the set $\{1, 2, \dots, n\}$. For each index i in this subset, pick a connected component of ∂K_i . Finally, applying Theorem 1 and/or Theorem 2 we can find a set of points (s_1, \dots, s_m) which is $(m-1)$ -dimensional in \mathbb{R}^m . We do the same for all possible choices of $\ell, 1 \leq \ell \leq n$, of the indices i_1, \dots, i_ℓ from the set $\{1, \dots, n\}$ and of connected components of $\partial K_{i_1}, \dots, \partial K_{i_\ell}$, respectively, and for $\ell = 0$. We denote the union of all the $(m-1)$ -dimensional sets, which are included in V , by L_0 .

Next, consider a finite set of points, say b_1, \dots, b_p , in $\bigcup_{i=1}^n \partial K_i$ which correspond to the points where the derivatives of $A_i(\theta_i)$ do not exist. In the sequel, we will label such points "bad" points. Suppose that $b_1 \in \partial K_1$. By substituting $A_1 = b_1$, $f(A, s) = 0$ reduces to a (vector) equation in the (complex) unknowns A_2, \dots, A_n and $s, s \in \mathbb{R}^m$. We apply Theorem 1 and/or Theorem 2 to the new equation at the point (b_1, A_2, \dots, A_n) and obtain $(m-1)$ -dimensional sets in \mathbb{R}^m , included in V , in a way similar to the previous procedure for L_0 . Next we repeat the procedure for b_2, b_3 , etc. up to b_p . Denote by L_1 the union of all these $(m-1)$ -dimensional sets for b_1, \dots, b_p .

Next, we substitute in $f(A, s) = 0$ two bad points, say b_{n_1} for A_i and b_{n_2} for A_k , such that $i \neq k$, and apply Theorem 1 and/or Theorem 2. We denote by L_2 the union of all the sets, which are included in V . Next, we choose three bad points, then four bad points, etc., and obtain L_3, \dots, L_p . Let

$$L = \bigcup_{i=0}^p L_i.$$

Then, since each L_i is $(m-1)$ -dimensional and included in V , so is L and

$$\partial_{\ell_i} V \subset L \subset V,$$

where $\partial_{\ell_i} V = \partial V \cap G$.

3. GENERALIZED ZERO SETS LOCATION

A. The generalized zero set

Let $Q = Q_1 \times \dots \times Q_n$ and $P = P_1 \times \dots \times P_k$ be sets in $\mathbb{C}^n = \mathbb{C} \times \dots \times \mathbb{C}$ (n times) and $\mathbb{C}^k = \mathbb{C} \times \dots \times \mathbb{C}$ (k times), respectively, where each Q_i and P_i is a closed set in $\mathbb{C} = \mathbb{C} \cup \{\infty\}$. For each i the boundary ∂Q_i and ∂P_i is a finite union of piecewise-smooth simple curves and piecewise-smooth closed simple curves. Let G be an open set in \mathbb{R}^m and let $f: Q \times P \times G \rightarrow \mathbb{R}^d \cup \{\infty\}$ be a continuously differentiable function.

The last assumption means that $Q \times P \times G$ has an open neighborhood, where $f(q, p, s) = (f_1(q, p, s), \dots, f_d(q, p, s))$ has continuous partial derivatives with respect to all real coordinates $q_i^1, q_i^2, p_i^1, p_i^2$, and s_i , where $q_i = q_i^1 + jq_i^2$, $p_i = p_i^1 + jp_i^2$, $q = (q_1, \dots, q_n) \in Q$, $p = (p_1, \dots, p_k) \in P$ and $s = (s_1, \dots, s_m) \in G$. The generalized zero set of $f = (f_1, \dots, f_d)$ relative to Q, P and G is then defined by

$$M = \{s \in G \subset \mathbb{R}^m : \forall p \in P, \exists q \in Q \text{ such that } f(q, p, s) = 0\}. \quad (1)$$

In other words, $s \in M$ if and only if $f(p, q, s) = 0$ for some point q in Q and every point p in P .

Our main purpose in this section is to derive an algorithm which will enable us to determine and describe the generalized zero set M of the vector-valued function f , for certain parameter spaces Q and P , and open sets G . This objective will be carried out by redefining M in terms of certain zero sets.

B. The procedure of determining the generalized zero set

Let Q, P, G, f and M be as in Section 3A. In particular it is assumed that $Q \times P \times G$ has a neighborhood where f is continuously differentiable. Suppose that $\tilde{P} \supset P$ is an open set in \mathbb{C}^k for which f is continuously differentiable in $Q \times \tilde{P} \times G$. Let \tilde{V} be the zero set of f relative to Q and $G \times \tilde{P}$, i.e.

$$\tilde{V} = \{(s, p) \in G \times \tilde{P} \subset \mathbb{R}^m \times \mathbb{C}^k : \exists q \in Q \text{ such that } f(q, p, s) = 0\}. \quad (2)$$

Excluding points in which one of the coordinates of $p = (p_1, \dots, p_k) \in \tilde{P}$ is ∞ , we denote points $(s, p) \in G \times \tilde{P}$ by

$$(s, p) = (s_1, \dots, s_m, p_1^1, p_1^2, \dots, p_k^1, p_k^2),$$

where $(p_i^1 + jp_i^2) = p_i \in P_i \setminus \{\infty\}$.

Note that the differentiability of f at points (q, p, s) where at least one of the coordinates of (q, p) is ∞ or at points (q, p, s) for which $f(q, p, s) = \infty$, is checked by means of a change of variables of the form $z \rightarrow 1/z$, $z \in \mathbb{C}$.

Lemma 1. \tilde{V} is a closed set relative to $G \times \tilde{P}$. The proof is as in Fruchter *et al.* (1991a, Lemma IIA), where the idea is to show that \tilde{V} contains each of its limit points which are in $G \times \tilde{P}$. This follows from the compactness of Q (being a closed subset of \mathbb{C}^n) and the continuity of f .

The set \tilde{V} has all the properties that the zero set V considered in Fruchter (1988) and Fruchter *et al.* (1987, 1991a, b), and presented in Section 2, has, including the fact that \tilde{V} is closed relative

to $G \times \tilde{P}$. On the boundary of \tilde{V} relative to $G \times \tilde{P}$ we have similar conditions to those stated in Theorems 1 and 2. Therefore, in order to find the location of \tilde{V} in $G \times \tilde{P}$ we may apply the method briefed in Section 2.

The following proposition is an immediate consequence of (1) and (2).

Proposition 1. Let Q, P, G, f, M and \tilde{V} be as in 3A. Then

$$M = \{s \in G : \forall p \in P, (s, p) \in \tilde{V}\}. \quad (3)$$

Now, let L be an $(m-1)$ -dimensional set, as described in Section 2, i.e. L is included in \tilde{V} and includes the boundary of the zero set \tilde{V} . Then L is a finite union of solutions of equations obtained from Theorem 1 and/or Theorem 2. In other words L , and consequently the relative boundary of \tilde{V} , is a finite union of $(m-1)$ -dimensional sets of the form

$$\{(s, p) \in G \times \tilde{P} : \varphi_i(p, s) = 0\},$$

where the equations $\varphi_i = 0$ are derived from the conditions on the relative boundary of \tilde{V} which Theorems 1 and 2 provide. From the implicit function theorem it follows that φ_i are continuously differentiable real valued functions. Therefore, in many cases, the zero set \tilde{V} can be expressed as a finite union and intersection of the sets

$$\tilde{V}_i = \{(s, p) \in G \times \tilde{P} : \varphi_i(p, s) \leq 0\}. \quad (4)$$

Using Proposition 1 we obtain, in this case, that M can be expressed by a corresponding finite union and intersection of the sets

$$M_i = \{s \in G : \forall p \in P, \varphi_i(p, s) \leq 0\}. \quad (5)$$

The complement of M_i in G , denoted by S_i , becomes

$$S_i = G \setminus M_i = \{s \in G : \exists p \in P, \varphi_i(p, s) > 0\}.$$

Note that

$$\varphi_i(p, s) > 0$$

is equivalent to

$$\varphi_i(p, s) + n = 0 \quad \text{for some } n \in [-\infty, 0).$$

Denote

$$h_i(n, p, s) = \varphi_i(p, s) + n. \quad (6)$$

Then S_i becomes

$$S_i = \{s \in G : \exists p \in P, \exists n \in [-\infty, 0) \text{ such that } h_i(n, p, s) = 0\}.$$

It is easy to see that

$$S_i = \bigcup_{\epsilon < 0} S_i', \quad (7)$$

where

$$S_i^* = \{s \in G: \exists p \in P, \exists n \in [-\infty, \infty] \text{ such that } h_i(n, p, s) = 0\}. \quad (8)$$

The set S_i^* is recognized as a zero set, presented in Section 2 of the real-valued function h_i . In conclusion the sets $M_i = G \setminus S_i$, and therefore M , can be found by the method presented in Section 2.

In conclusion, M can be found by the following steps.

Step 1: Choose \tilde{P} and determine the corresponding set \tilde{V} [see (2)].

Step 2: Write \tilde{V} as a finite union and intersection of sets \tilde{V}_i [see (4)].

Step 3: Write M as a finite union and intersection of sets M_i [see (5)].

Step 4: Determine the functions h_i from φ_i [see (6)] and write the complements S_i^* of M_i as a union of the zero sets S_i^* , on $t \leq 0$, [see (7) and (8)].

Step 5: Determine each S_i and M_i , and then M , by Step 3.

In the following we illustrate the procedure of finding M by two numerical examples.

C. Numerical examples

The following example is derived from a possible formulation of an absolute stability test of a certain nonlinear system by Jury-Lee criterion (Jury and Lee, 1964).

Example 1. Let $G = \{s = (s_1, s_2) \in \mathbb{R}^2: s_2 > 0\}$, let $Q = Q_1 \times Q_2$, where

$$Q_1 = \{q_1 \in \mathbb{C}: |q_1| = 1\}$$

$$Q_2 = \{q_2 \in \mathbb{C}: \operatorname{Re} q_2 \geq 0\} \cup \{\infty\},$$

and let $P = [0, \infty)$. Let $f: Q \times P \times G \rightarrow \mathbb{R}^2 \cup \{\infty\}$ be defined by

$$f(q_1, q_2, p, s_1, s_2) = p[(0.3 - 0.05s_1) + (0.1 - 0.03s_1) \operatorname{Re} q_1] - (0.2 + 0.1 \operatorname{Re} q_1) + 1/s_2 + q_2.$$

We want to find the following set:

$$\mathcal{M} = \{s \in G: \forall p \in P, \exists (q_1, q_2) \in Q \text{ such that } f(q_1, q_2, p, s_1, s_2) = 0\}.$$

Determination of the set \mathcal{M} . Let $P_\alpha = [0, \alpha]$. It is easy to see that

$$P = \bigcup_{\alpha > 0} P_\alpha.$$

Hence

$$\mathcal{M} = \bigcap_{\alpha > 0} M^\alpha, \quad (9)$$

where

$$M^\alpha = \{s \in G: \forall p \in P_\alpha, \exists (q_1, q_2) \in Q \text{ such that } f(q_1, q_2, p, s_1, s_2) = 0\}.$$

The set M^α is recognized as a generalized zero set of the complex-valued function f . In order to find M^α , and therefore \mathcal{M} , we apply the procedure presented above.

Step 1: We set $\tilde{P} = \mathbb{R}$ and find the set

$$\tilde{V} = \{(s_1, s_2, p) \in G \times \mathbb{R}: \exists (q_1, q_2) \in Q \text{ such that } f(q_1, q_2, p, s_1, s_2) = 0\}.$$

In Fruchter *et al.* (1987a, Example 2) we found this set and we obtained the following result:

$$\tilde{V} = \{(s_1, s_2, p) \in G \times \mathbb{R}: -0.3 + p(0.4 - 0.08s_1) + 1/s_2 \leq 0 \text{ or } -0.1 + p(0.2 - 0.02s_1) + 1/s_2 \leq 0\}.$$

Step 2: The set \tilde{V} can be written as a union of two sets \tilde{V}_1 and \tilde{V}_2 , namely

$$\tilde{V} = \tilde{V}_1 \cup \tilde{V}_2,$$

where

$$\begin{aligned} \tilde{V}_1 &= \{(s_1, s_2, p) \in G \times \mathbb{R}: \varphi_1(p, s_1, s_2) \\ &= -0.3 + p(0.4 - 0.08s_1) + 1/s_2 \leq 0\} \end{aligned}$$

and

$$\begin{aligned} \tilde{V}_2 &= \{(s_1, s_2, p) \in G \times \mathbb{R}: \varphi_2(p, s_1, s_2) \\ &= -0.1 + p(0.2 - 0.02s_1) + 1/s_2 \leq 0\}. \end{aligned}$$

Step 3: The set M^α , can be written as a union of two sets M_1^α and M_2^α , namely,

$$M^\alpha = M_1^\alpha \cup M_2^\alpha,$$

where

$$M_1^\alpha = \{(s_1, s_2) \in G: \forall p \in P_\alpha, \varphi_1(p, s_1, s_2) \leq 0\}$$

and

$$M_2^\alpha = \{(s_1, s_2) \in G: \forall p \in P_\alpha, \varphi_2(p, s_1, s_2) \leq 0\}.$$

Now, let $M_1 = \bigcap_{\alpha > 0} M_1^\alpha$ and $M_2 = \bigcap_{\alpha > 0} M_2^\alpha$, then from (9) follows that

$$\mathcal{M} = M_1 \cup M_2.$$

Step 4: The functions h_i , $i = 1, 2$ will have the form

$$h_1(n, p, s) = -0.3 + p(0.4 - 0.08s_1) + 1/s_2 + n$$

and

$$h_2(n, p, s) = -0.1 + p(0.2 - 0.02s_1) - 1/s_2 + n.$$

Now, the complement of M_i^α in G , denoted by $S_i^{\alpha*}$, will be

$$S_i^{\alpha*} = G \setminus M_i^\alpha = \bigcup_{p \in P_\alpha} S_i^{p*}, \quad i = 1, 2$$

where

$$S_i^{\varepsilon, \alpha} = \{(s_1, s_2) \in G: \exists p \in [0, \alpha], \exists n \in [-\infty, \varepsilon] \text{ such that } h_i(n, p, s) = 0\}.$$

The complement of M_i in G , denoted by S_i , is then given by

$$S_i = G \setminus M_i = \bigcup_{\alpha \rightarrow 0} \bigcup_{\varepsilon \rightarrow 0} S_i^{\varepsilon, \alpha}, \quad i = 1, 2$$

The sets S_i , $i = 1, 2$, are obtained from $S_i^{\varepsilon, \alpha}$ by taking $\varepsilon \rightarrow 0^+$ and $\alpha \rightarrow +\infty$.

Step 5: Determination of $S_i^{\varepsilon, \alpha}$ and S_i , $i = 1, 2$. The sets $S_i^{\varepsilon, \alpha}$, $i = 1, 2$, are zero sets of the continuously differentiable real-valued functions h_i , respectively, relative to $[-\infty, \varepsilon] \times [0, \alpha]$ and G . For finding $S_i^{\varepsilon, \alpha}$ we use the procedure briefed in Section 2. As was mentioned above S_i is obtained from $S_i^{\varepsilon, \alpha}$ by taking $\varepsilon \rightarrow 0^+$ and $\alpha \rightarrow +\infty$.

As outlined in this procedure, first one finds the set L . In the present example $L = L_0 \cup L_1 \cup L_2$. For finding L_j , $j = 0, 1, 2$, we apply Theorem 1, with $d = 1$ (see Section 2).

Since

$$\frac{\partial h_i}{\partial n} = 1 \neq 0, \quad i = 1, 2, \quad (10)$$

we obtain immediately by Theorem 1 that

$$L_0 = \emptyset.$$

Also, it is readily verified that

$$h_i(-\infty, p, s_1, s_2) \neq 0, \quad i = 1, 2. \quad (11)$$

Therefore, from (10) and (11) we obtain that in the derivation of L , we have to consider only the following cases.

In the derivation of L_1 we have only the case:

$$h_i(\varepsilon, p, s_1, s_2) = 0, \quad 0 < p < \alpha, \quad i = 1, 2 \quad (12)$$

and for L_2 we have the cases:

$$h_i(\varepsilon, 0, s_1, s_2) = 0, \quad i = 1, 2 \quad (13)$$

$$h_i(\varepsilon, \alpha, s_1, s_2) = 0, \quad i = 1, 2. \quad (14)$$

First, we treat the case $i = 1$.

The equations of Theorem 1 which correspond to the case (12) are

$$h_1(\varepsilon, p, s_1, s_2) = -0.3 + p(0.4 - 0.08s_1) + 1/s_2 + \varepsilon = 0 \quad (15)$$

$$\frac{\partial h_1}{\partial p} = 0.4 - 0.08s_1 = 0. \quad (16)$$

From (15) and (16), taking $\varepsilon \rightarrow 0^+$, we obtain

$$L_1 = \{(5, 10/3)\}.$$

Now, the equations of Theorem 1 for the case

(13) are reduced to

$$h_1(\varepsilon, 0, s_1, s_2) = -0.3 + 1/s_2 + \varepsilon = 0.$$

Taking $\varepsilon \rightarrow 0^+$, we obtain

$$s_2 = 10/3. \quad (17)$$

The equations of Theorem 1 for the case (14) are reduced to

$$h_1(\varepsilon, \alpha, s_1, s_2) = -0.3 + \alpha(0.4 - 0.08s_1) + 1/s_2 + \varepsilon = 0.$$

Taking $\varepsilon \rightarrow 0^+$, we obtain

$$5 - s_1 = \frac{3(s_2 - 10/3)}{0.8\alpha s_2}.$$

Hence, taking $\alpha \rightarrow +\infty$, if $s_2 \neq 10/3$, we obtain

$$s_1 = 5. \quad (18)$$

Therefore, from (17) and (18) we obtain

$$L_2 = \{(s_1, s_2) \in G: s_1 = 5 \text{ or } s_2 = 10/3\}.$$

And in conclusion,

$$L = \{(s_1, s_2) \in G: s_1 = 5 \text{ or } s_2 = 10/3\}.$$

The set L which is a one-dimensional set, is depicted in Fig. 1 and divides G into four domains D_i , $i = 1, \dots, 4$. In order to decide which of the domains D_i belongs to S_1 , we choose arbitrary point in each of the domains D_i and check whether these points belong to S_1 . It is readily verified that

$$S_1 = \bigcup_{i=1}^4 D_i.$$

The complement of S_1 in G is dashed in Fig. 1 and is given by

$$M_1 = \bar{D}_4 = \{(s_1, s_2) \in G: s_1 \geq 5 \text{ and } s_2 \geq 10/3\}.$$

For the case $i = 2$ we have similar computations. We obtained that $S_1 \subset S_2$, therefore $M_1 \supset M_2$. Hence

$$\mathcal{M} = M_1 \cup M_2 = M_1 = \bar{D}_4.$$

And in conclusion,

$$\mathcal{M} = \{(s_1, s_2) \in G: s_1 \geq 5 \text{ and } s_2 \geq 10/3\},$$

namely the dashed region in Fig. 1.

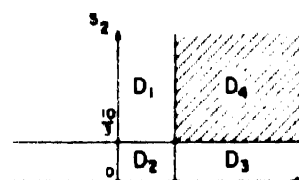


FIG. 1. The generalized zero set $\mathcal{M} = \bar{D}_4$ for Example 1 and the connected components of $G \setminus L$.

Remark 1. If P and M are subsets of \mathbb{R} then \bar{V} is a subset of \mathbb{R}^2 [see (2) Section 3B] and M can be found directly from the geometrical interpretation of \bar{V} . This is an immediate consequence of Proposition 1(3). In such cases we are able to find M immediately after Step 1, as it will be illustrated in the following example. In consequence, in this case, we do not need the closedness of the set P .

Example 2. Let $G = \{s \in \mathbb{R}^+ : s > 0\}$, let $Q = Q_1 \times Q_2$, where $Q_1 = Q_2 = [0, \infty)$ and let $P = \mathbb{R}$. Let $f: Q \times P \times G \rightarrow \mathbb{R} \cup \{\infty\}$ be defined by

$$f(q_1, q_2, p, s) = s^{-1} - 6/(36q_1^2 + (5 - q_1^2)^2) + p(5 - q_1^2)/(36q_1^2 + (5 - q_1^2)^2) + q_2.$$

We want to find the following set:

$$\mathcal{M} = \{s \in G : \forall p \in P, \exists (q_1, q_2) \in Q \text{ such that } f(q_1, q_2, p, s) = 0\}.$$

Determination of the set \mathcal{M} . Let $Q_1^\alpha = [0, \alpha]$ and $Q_2^\beta = [0, \beta]$. Then, $Q_1 = \bigcup_{\alpha \rightarrow +\infty} Q_1^\alpha$ and $Q_2 = \bigcup_{\beta \rightarrow +\infty} Q_2^\beta$. Hence

$$\mathcal{M} = \bigcup_{\substack{\alpha \rightarrow +\infty \\ \beta \rightarrow +\infty}} M^{\alpha, \beta}$$

where

$$M^{\alpha, \beta} = \{s \in G : \forall p \in P, \exists (q_1, q_2) \in Q_1^\alpha \times Q_2^\beta \text{ such that } f(q_1, q_2, p, s) = 0\}.$$

According to Remark 1, the set $M^{\alpha, \beta}$ is recognized as a generalized zero set of the real-valued function f . In order to find $M^{\alpha, \beta}$, and therefore \mathcal{M} , we need to use only the Step 1 of the procedure presented above.

Step 1: We have $\bar{P} = P = \mathbb{R}$. Let

$$\bar{V}^{\alpha, \beta} = \{(s, p) \in G \times \mathbb{R} : \exists (q_1, q_2) \in Q_1^\alpha \times Q_2^\beta \text{ such that } f(q_1, q_2, p, s) = 0\}.$$

Then

$$\bar{V} = \bigcup_{\substack{\alpha \rightarrow +\infty \\ \beta \rightarrow +\infty}} \bar{V}^{\alpha, \beta}.$$

The set $\bar{V}^{\alpha, \beta}$ is recognized as a zero set of the continuously differentiable real-valued function f relative to $Q_1^\alpha \times Q_2^\beta = [0, \alpha] \times [0, \beta]$ and $G \times \mathbb{R}$. Therefore, for finding $\bar{V}^{\alpha, \beta}$, we use the algorithm briefed in Section 2. From $\bar{V}^{\alpha, \beta}$, by taking $\alpha \rightarrow +\infty$ and $\beta \rightarrow +\infty$, we obtain \bar{V} .

Determination of $\bar{V}^{\alpha, \beta}$ and \bar{V} . As outlined in the procedure, first one finds the set L . In the present example $L = L_0 \cup L_1 \cup L_2$. For finding L_j , $j = 0, 1, 2$, we apply Theorem 1, with $d = 1$. (See Section 2.)

Since

$$\frac{\partial f}{\partial q_2} = 1 \neq 0 \quad (18)$$

we obtain immediately by Theorem 1 that

$$L_0 = \emptyset.$$

Therefore, in the derivation of L , we have to consider the following cases:

In the derivation of L_1 we have the cases

$$f(q_1, 0, p, s) = 0, \quad 0 < q_1 < \alpha \quad (19a)$$

$$f(q_1, \beta, p, s) = 0, \quad 0 < q_1 < \alpha \quad (19b)$$

and for L_2 we have the cases

$$f(0, 0, p, s) = 0 \quad (20a)$$

$$f(0, \beta, p, s) = 0 \quad (20b)$$

$$f(\alpha, 0, p, s) = 0 \quad (20c)$$

$$f(\alpha, \beta, p, s) = 0. \quad (20d)$$

It is readily verified that when $\alpha \rightarrow +\infty$, $\beta \rightarrow +\infty$ and $s > 0$, only (19a) and (20a) are meaningful.

Now, in the case (19a) the equation

$$f(q_1, 0, p, s) = s^{-1} - 6/(36q_1^2 + (5 - q_1^2)^2) + p(5 - q_1^2)/(36q_1^2 + (5 - q_1^2)^2) = 0$$

is equivalent to the equation

$$f^*(q_1^2, p, s) = s^{-1}(36q_1^2 + (5 - q_1^2)^2) - 6 + p(5 - q_1^2) = 0. \quad (21)$$

Hence, we obtain by Theorem 1 that, in case (19a), we have in addition to (21) the equation

$$\frac{\partial f^*}{\partial q_1^2} = s^{-1}(36 - 2(5 - q_1^2)) - p = 0. \quad (22)$$

From (21) and (22) we obtain the solution

$$p^2s^2 - 72ps + 576 + 24s = 0. \quad (23)$$

In the case (20a) we obtain

$$f(0, 0, p = a) = s^{-1} - 6/25 + p/5 = 0$$

or

$$p = 6/5 - 5s^{-1}. \quad (24)$$

From (23) and (24) we obtain

$$L = \{(s, p) \in G \times \mathbb{R} : p^2s^2 - 72ps + 576 + 24s = 0 \text{ or } p = 6/5 - 5s^{-1}\}.$$

The set L which is a one-dimensional set, is depicted in Fig. 2 and divides $G \times \mathbb{R}$ into four connected domains D_i , $i = 1, \dots, 4$. In order to decide which of the domains D_i belongs to \bar{V} , we choose arbitrary points in each of the domains D_i and check whether these points belong to \bar{V} . It is readily verified that

$$\bar{V} = \bigcup_{i=1}^4 D_i \cup L.$$

The set \bar{V} is dashed in Fig. 2. From this sketch and Proposition 1 we can conclude immediately

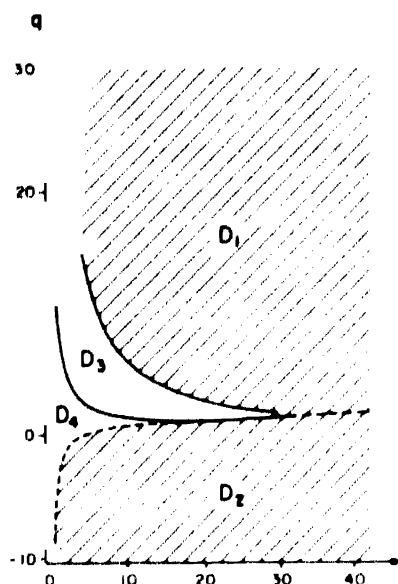


FIG. 2. The zero set V for Example 2 and the connected components of $\mathbb{R}^n \times \mathbb{R} \setminus L$.

that

$$\mathcal{M} = [30, +\infty).$$

Remark. The complement of \mathcal{M} in G has an engineering meaning which will be explained in Section 4 (Remark 4).

4. ABSOLUTE ROBUST STABILIZATION OF CONTINUOUS NONLINEAR CONTROL SYSTEMS

The nature of the mathematical results of the previous section seems to be particularly pertinent to various applications in engineering system theory. In this section, we apply the results to one such problem, which is important from viewpoint of practical design of systems under uncertainty conditions and which is considered very difficult to carry out. We consider the problem of robustness of absolute stability of Lurie type (Lurie, 1954) continuous nonlinear systems, with linear parts under uncertainty conditions.

Lurie-Postnikov's (Lurie and Postnikov, 1944) concept of the absolute stability of a class of sector nonlinear (so-called Lurie) systems can be considered as the concept of the robust global asymptotic stability with respect to variations, uncertainties and impreciseness of the *nonlinearity*. Absolute stability of a Lurie system is tested in the literature for a given (nominal, unperturbed) mathematical description of the system. However, it is of utmost practical importance to consider also the uncertainties and impreciseness of the parameters of the linear part of the system. The robust absolute stability and stabilization for a continuous sector nonlinear system with linear part parameter

uncertainties, is studied in this section. Applying the method of "generalized zero sets location", we find all the feasible set of sectors of nonlinearities, for which absolute robust stability with respect to the uncertainty conditions is ensured, according to the Popov criterion (Popov, 1961).

A. System description

A nominal description of a Lurie continuous-time system with one nonlinear unit ϕ is defined by

$$\dot{x} = Ax + b\phi(\sigma), \quad x = x(t) \in \mathbb{R}^n, \\ A \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^{n \times 1}, \quad \sigma \in \mathbb{R} \quad (25a)$$

$$\sigma = cx, \quad c \in \mathbb{R}^1 \quad (25b)$$

The nonlinearity ϕ (Lurie nonlinearity) is a continuous function from \mathbb{R} to \mathbb{R} satisfying the following sector condition:

$$\phi(0) = 0 \quad \text{and} \quad 0 < \phi(\sigma)\sigma^{-1} < k \quad \text{for} \quad \sigma \neq 0,$$

for a given positive number k , called sector number. Note that the above condition restricts the graph of the nonlinearity ϕ to within specified sector $S(k)$ which lies in the first and third quadrants of the plane and is bounded by the σ -axis and the line $y = k\sigma$.

Assume that some or all the entries in A , b , c are subject to perturbations and are only known to be within given intervals.

A perturbed model of the Lurie continuous system with one nonlinear unit is given by

$$\dot{x} = A(r)x + b(r)\phi(\sigma) \quad (26a)$$

$$\sigma = c(r)x, \quad (26b)$$

where $r = (r_1, \dots, r_q)$, the vector of all uncertainties of the system, may take values in $\prod_{l=1}^q K_l$, where $r_l \in K_l$ and $K_l = [\alpha_l, \beta_l]$, $l = 1, \dots, q$, are given intervals in \mathbb{R} .

B. Problem statement

Our purpose is to find the complete set of sector numbers k for which the system (26) is absolutely robustly stable for every $r \in \prod_{l=1}^q K_l$, and for every $\phi(\sigma)$ which satisfies the sector condition, according to Popov criterion (Popov, 1961). The significance of the solution to this problem, relative to previous works, was explained in the introduction.

C. Problem solution via generalized zero sets location

We will assume that $A(r)$ of the system (26) is a robust stable matrix. If this is not the case we

first stabilize $A(r)$ robustly by a constant output feedback h as in (Fruchter *et al.*, 1987a, 1991a) and replace $A(r)$ by

$$A^*(r) = A(r) + b(r)hc(r).$$

We consider now the design problem of finding the sector numbers k for which the system (26) is absolutely robustly stable. For this purpose we generalize a criterion by Popov (1961) and apply our generalized zero set method.

Let A , b , c , σ and $\phi(\sigma)$ be as in (25). Let $G_1(\lambda) = c(A - \lambda I)^{-1}b$, where λ is the complex variable, be the transfer function of the linear part of (25), from the input $\phi(\sigma)$ to the output $-\sigma$. Assuming that A is a stable matrix, that is, all the zeros of the polynomial

$$\Delta_1(\lambda) = |(A - \lambda I)|$$

lie in the open left half complex plane $\text{Re } \lambda < 0$, and assuming that the rational function $G_1(\lambda)$ has no cancelable factors, Popov (1961) proved the following theorem.

Theorem 3 (Popov, 1961). The system in (25) is absolutely stable if there exist $q \in \mathbb{R}$ such that

$$k^{-1} + \text{Re}[(1 + j\omega q)G_1(j\omega)] > 0$$

for all $\omega \geq 0$.

Now, let $G_2(\lambda, r) = c(r)(A(r) - \lambda I)^{-1}b(r)$ be the transfer function of the linear part of (26), from the input $\phi(\sigma)$ to the output $-\sigma$, and assume that $A(r)$ is a robust stable matrix, for all $r \in \prod_{l=1}^q K_l$. An immediate consequence of Theorem 3 is the following.

Corollary 1. The system in (26) is absolutely robust stable if there exist $q \in \mathbb{R}$ such that

$$k^{-1} + \text{Re}[(1 + j\omega q)G_2(j\omega, r)] > 0$$

for all $\omega \geq 0$ and all $r \in \prod_{l=1}^q K_l$.

Let

$$\begin{aligned} \varphi(\omega, r, q, k) &= k^{-1} + \text{Re}[(1 + j\omega q)G_2(j\omega, r)] \\ &= k^{-1} + \text{Re } G_2(j\omega, r) \\ &\quad - q\omega \text{Im } G_2(j\omega, r). \end{aligned}$$

Then the objective is to determine the set S_{Popov} of all points $k \in \mathbb{R}^+$, where

$$\mathbb{R}^+ = \{k \in \mathbb{R} : k > 0\},$$

for which (26) is absolutely robustly stable. By

Corollary 1 S_{Popov} has the form

$$S_{\text{Popov}} = \left\{ k \in \mathbb{R}^+ : \exists q \in \mathbb{R}, \text{ such that} \right.$$

$$\left. \varphi(\omega, r, q, k) > 0, \forall \omega \geq 0 \text{ and } \forall r \in \prod_{l=1}^q K_l \right\}.$$

To adopt the above formulation to the generalized set location method, we use an auxiliary variable p , $p \in [0, \infty]$, in the following way: Let

$$f(\omega, r, p, q, k) = \varphi(\omega, r, q, k) + p.$$

Then

$$\varphi(\omega, r, q, k) > 0$$

is equivalent to

$$\begin{aligned} f(\omega, r, p, q, k) &= \varphi(\omega, r, q, k) \\ &\quad + p \neq 0, \forall p \in [0, \infty] \end{aligned}$$

and the set S_{Popov} becomes

$$\begin{aligned} S_{\text{Popov}} &= \left\{ k \in \mathbb{R}^+ : \exists q \in \mathbb{R} \right. \\ &\quad \left. \text{such that } f(\omega, r, p, q, k) \neq 0, \right. \\ &\quad \left. \forall \omega \geq 0, \forall r \in \prod_{l=1}^q K_l \text{ and } \forall p \in [0, \infty] \right\}. \quad (27) \end{aligned}$$

Remark 2. Let $\omega \geq 0$, $r \in \prod_{l=1}^q K_l$, $p \in [0, \infty]$, $q \in \mathbb{R}$ and $k \in \mathbb{R}^+$. For stable matrices $A(r)$, we obtain continuously differentiable functions $\varphi(\omega, r, q, k)$, and therefore, $f(\omega, r, p, q, k)$.

Remark 3. For each point k^0 in S_{Popov} , our system is robust stable with respect to any r in $\prod_{l=1}^q K_l$ and any nonlinearity ϕ in the sector $S(k^0)$.

Remark 4. The complement of the set \mathcal{M} in Example 2 is exactly the set S_{Popov} for an unperturbed system.

From (27), we obtain that the complement of S_{Popov} in \mathbb{R}^+ will be

$$\mathcal{M}_{\text{Popov}} = \mathbb{R}^+ \setminus S_{\text{Popov}} = \left\{ k \in \mathbb{R}^+ : \forall q \in \mathbb{R}, \exists \omega \geq 0, \right.$$

$$\left. \exists r \in \prod_{l=1}^q K_l, \exists p \in [0, \infty] \right.$$

$$\left. \text{such that } f(\omega, r, p, q, k) = 0 \right\}.$$

It is easy to see that

$$\mathcal{M}_{\text{Popov}} = \bigcup_{\alpha > 0} \mathcal{M}^\alpha.$$

where

$$M^{\alpha} = \left\{ k \in \mathbb{R}^+ : \forall q \in \mathbb{R}, \exists \omega \in [0, \alpha], \right. \\ \left. \exists r \in \prod_{l=1}^q K_l, \exists p \in [0, \infty] \right. \\ \left. \text{such that } f(\omega, r, p, q, k) = 0 \right\}.$$

The set M^{α} is recognized as a generalized zero set of the continuously differentiable real-valued function f relative to $Q = [0, \alpha] \times \left(\prod_{l=1}^q K_l \right) \times [0, \infty]$, $P = \mathbb{R}$ and $G = \mathbb{R}^+$. Therefore, for finding M^{α} we may apply the procedure presented in Section 3. Since $P = \mathbb{R}$ and M^{α} is a subset of \mathbb{R} , according to Remark 2 we need only the Step 1 of this procedure. Let

$$\tilde{V}^{\alpha} = \left\{ (k, q) \in \mathbb{R}^+ \times \mathbb{R} : \exists (\omega, r, p) \in [0, \alpha] \right. \\ \left. \times \left(\prod_{l=1}^q K_l \right) \times [0, \infty] \right. \\ \left. \text{such that } f(\omega, r, p, q, k) = 0 \right\}$$

be the zero set of f relative to $[0, \alpha] \times \left(\prod_{l=1}^q K_l \right) \times [0, \infty]$ and $\mathbb{R}^+ \times \mathbb{R}$. Then by Proposition 1 we obtain

$$M^{\alpha} = \{ k \in \mathbb{R}^+ : \forall q \in \mathbb{R}, (k, q) \in \tilde{V}^{\alpha} \},$$

and hence

$$\mathcal{M}_{\text{Popov}} = \{ k \in \mathbb{R}^+ : \forall q \in \mathbb{R}, (k, q) \in \tilde{V} \}, \quad (28)$$

where

$$\tilde{V} = \bigcup_{\alpha=0} \tilde{V}^{\alpha}.$$

The set \tilde{V} can be obtained from \tilde{V}^{α} by taking $\alpha \rightarrow +\infty$.

In conclusion, the problem of absolute robust stabilization of a Lurie type continuous-time system with one nonlinearity unit, according to the Popov criterion, is reduced to the problem of locating a generalized zero set in the open subset \mathbb{R}^+ of \mathbb{R} . This set can be determined immediately after Step 1 of the algorithm presented in the previous section. In this way, the complete set of sector numbers k can always (for any numbers of uncertainty parameters in the linear part, and all forms of their appearance) be selected from a two-dimensional

set:

$$\mathcal{P} = \mathbb{R}^+ \times \mathbb{R} \setminus V = \left\{ (k, q) \in \mathbb{R}^+ \times \mathbb{R} : k^{-1} \right. \\ \left. + \operatorname{Re} [(1 + j\omega q)G(j\omega, r)] > 0, \right. \\ \left. \forall \omega \geq 0, \forall r \in \prod_{l=1}^q K_l \right\}.$$

In the following we illustrate the procedure of finding $\mathcal{M}_{\text{Popov}}$ by a numerical example.

D. Numerical example

Consider the continuous system with uncertainty, described in (26), with the transfer function $G_2(\lambda, r)$ given by

$$G_2(\lambda, r) = \frac{1}{\lambda(1 + \lambda)(1 + r\lambda + \lambda^2/16)}. \quad (29)$$

The uncertainty parameter r assumes values in the interval $[0.001, 1/2]$. The nominal value of r is $0.4/16$.

Let S_{Popov} be the set of points k in \mathbb{R}^+ for which this system is absolutely robust stable with respect to any variation of the nonlinearities ϕ in the sector $S(k)$ and of r in the interval $[0.001, 1/2]$. By the above discussion we obtain that

$$S_{\text{Popov}} = \{ k \in \mathbb{R}^+ : \exists q \in \mathbb{R} \text{ such that } \\ f(\omega, r, p, q, k) \neq 0, \forall \omega \geq 0, \\ \forall r \in [0.001, 1/2] \text{ and } \forall p \in [0, \infty] \},$$

where

$$f(\omega, r, p, q, k) = k^{-1} + \operatorname{Re} G_2(j\omega, r) \\ - q\omega \operatorname{Im} G_2(j\omega, r) + p,$$

and $G_2(j\omega, r)$ is given by (29), namely

$$G_2(j\omega, r) = \frac{(1 - \omega^2/16 + r)}{(1 + \omega^2)((1 - \omega^2/16)^2 + r^2\omega^2)} \\ + j \frac{(1 - \omega^2/16 - r\omega^2)}{\omega(1 + \omega^2)((1 - \omega^2/16)^2 + r^2\omega^2)}.$$

Evidently, for r in $[0.001, 1/2]$, $G_2(\lambda, r)$ has no poles in $\operatorname{Re} \lambda \geq 0$, and in conclusion $f(\omega, r, p, q, k)$ is continuously differentiable (Remark 2).

The complement of S_{Popov} in \mathbb{R}^+ , denoted by $\mathcal{M}_{\text{Popov}}$, becomes $\mathcal{M}_{\text{Popov}} = \mathbb{R}^+ \setminus S_{\text{Popov}} = \{ k \in \mathbb{R}^+ : \forall q \in \mathbb{R}, \exists \omega \geq 0,$

$$\exists r \in [0.001, 1/2], \exists p \in [0, \infty] \\ \text{such that } f(\omega, r, p, q, k) = 0 \}.$$

Setting

$$\mathcal{M}_{\text{Popov}}^{\alpha} = \{ k \in \mathbb{R}^+ : \forall q \in \mathbb{R}, \exists \omega \in [0, \alpha], \\ \exists r \in [0.001, 1/2], \exists p \in [0, \infty] \\ \text{such that } f(\omega, r, p, q, k) = 0 \},$$

we obtain

$$\mathcal{M}_{\text{Popov}} = \bigcup_{\alpha > 0} M_{\text{Popov}}^{\alpha}.$$

The set $M_{\text{Popov}}^{\alpha}$ is the generalized zero set of the continuously differentiable function f , relative to $Q = [0, \alpha] \times [0.001, 1/2] \times [0, \infty]$, $P = \mathbb{R}$ and $G = \mathbb{R}^+$. Following the discussion in Section 3C, we define

$$\begin{aligned} \tilde{V}^{\alpha} = \{ (k, q) \in \mathbb{R}^+ \times \mathbb{R} : \exists (\omega, r, p) \in [0, \alpha] \\ \times [0.001, 1/2] \times [0, \infty] \\ \text{such that } f(\omega, r, p, q, k) = 0 \}. \end{aligned}$$

For finding \tilde{V}^{α} we apply results obtained in (Fruchter *et al.*, 1991a) and briefly summarized in Section 2.

As outlined in the procedure, first one finds the set L . In the present example $L = L_0 \cup L_1 \cup L_2 \cup L_3$. For finding L_j , $j = 0, 1, 2, 3$, we apply Theorem 1, with $d = 1$.

Since

$$\frac{\partial f}{\partial p} = 1 \neq 0 \quad (30)$$

we obtain immediately by Theorem 1 that

$$L_0 = \emptyset.$$

Also, it is readily verified that

$$f(\omega, r, +\infty, q, k) \neq 0. \quad (31)$$

Therefore, from (30) and (31) we obtain that in the derivation of L , we have to consider only the following cases:

In the derivation of L_1 we have the case

$$f(\omega, r, 0, q, k) = 0, \quad \omega \in (0, \alpha), \quad r \in (0.001, 1/2). \quad (32)$$

In the derivation of L_2 we have the cases

$$f(\omega, 0.001, 0, q, k) = 0, \quad \omega \in (0, \alpha) \quad (33)$$

$$f(\omega, 1/2, 0, q, k) = 0, \quad \omega \in (0, \alpha) \quad (34)$$

$$f(0, r, 0, q, k) = 0, \quad r \in (0.001, 1/2) \quad (35)$$

$$f(\alpha, r, 0, q, k) = 0, \quad r \in (0.001, 1/2). \quad (36)$$

Finally, in the derivation of L_3 we have the cases

$$f(0, 0.001, 0, q, k) = 0 \quad (37)$$

$$f(0, 1/2, 0, q, k) = 0 \quad (38)$$

$$f(\alpha, 0.001, 0, q, k) = 0 \quad (39)$$

$$f(\alpha, 1/2, 0, q, k) = 0. \quad (40)$$

It is readily verified that when $\alpha \rightarrow +\infty$, and $k > 0$, only (32)–(35) and (37)–(38) are meaningful.

Lets consider the case (32). It is easy to see

that the equation

$$\begin{aligned} f(\omega, r, 0, q, k) \\ = k^{-1} - \frac{(1 - \omega^2/16 + r)}{(1 + \omega^2)((1 - \omega^2/16)^2 + r^2\omega^2)} \\ + q \frac{(1 - \omega^2/16 - r\omega^2)}{(1 + \omega^2)((1 - \omega^2/16)^2 + r^2\omega^2)} = 0, \end{aligned}$$

is equivalent to the equation

$$\begin{aligned} f^*(\omega, r, q, k) = k^{-1}(1 + \omega^2)((1 - \omega^2/16)^2 + r^2\omega^2) \\ - (1 - \omega^2/16 + r) \\ - q(1 - \omega^2/16 - r\omega^2) = 0. \quad (41) \end{aligned}$$

Hence, we obtain by Theorem 1 that, in case (32), we have in addition to (41) the equations

$$\begin{aligned} \frac{\partial f^*}{\partial \omega^2} = k^{-1}((1 - \omega^2/16)^2 + r^2\omega^2) + (1 + \omega^2) \\ \times (-\frac{1}{k}(1 - \omega^2/16) + r^2) \\ + \frac{1}{16} - q(\frac{1}{16} + r) = 0 \quad (42) \end{aligned}$$

$$\frac{\partial f^*}{\partial r} = 2\omega^2rk^{-1}(1 + \omega^2) - 1 - q\omega^2 = 0. \quad (43)$$

The conditions (31), (41)–(43), yield a set of points $(k, q) \in \mathbb{R}^+ \times \mathbb{R}$ denoted by (44) in Figs 3a, b.

Applying Theorem 1 to the cases (33) and (34) we obtain the equations (41) and (42) with $r = 0.001$ and $r = 1/2$, respectively. These equations yield sets denoted, in Fig. 3, by (45) and (46), respectively.

Applying Theorem 1 to the case (35) we obtain the equations (41) and (43) with $\omega = 0$. These equation yield an empty set.

Finally, for the cases (37) and (38) we obtain

$$f^*(0, 0.001, q, k) = k^{-1} - 1.001 + q = 0 \quad (47)$$

and

$$f^*(0, 1/2, q, k) = k^{-1} - 3/2 + q = 0, \quad (48)$$

respectively.

In conclusion, the set L is given by the union of (44)–(48). The set L which is a one-dimensional set, is depicted in Fig. 3 and divides $\mathbb{R}^+ \times \mathbb{R}$ into 15 connected domains. In order to decide which of these domains belong to \tilde{V} , we choose arbitrary points in each of these domains and check whether these points belong to \tilde{V} . It is readily verified that

$$\tilde{V} = \mathbb{R}^+ \times \mathbb{R} \setminus D,$$

where D is the region dashed in Fig. 3. Hence, using (28) we obtain

$$\mathcal{M}_{\text{Popov}} = [0.6, +\infty).$$

Therefore, if $k \in S_{\text{Popov}}$, then, the considered

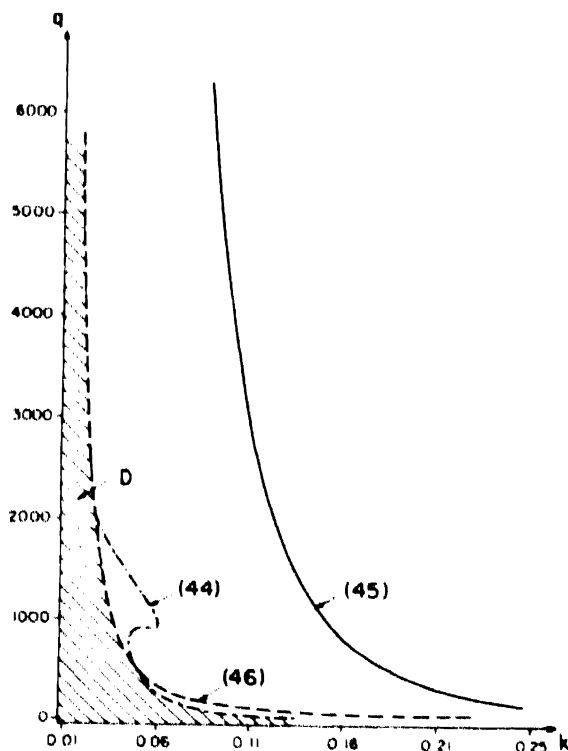
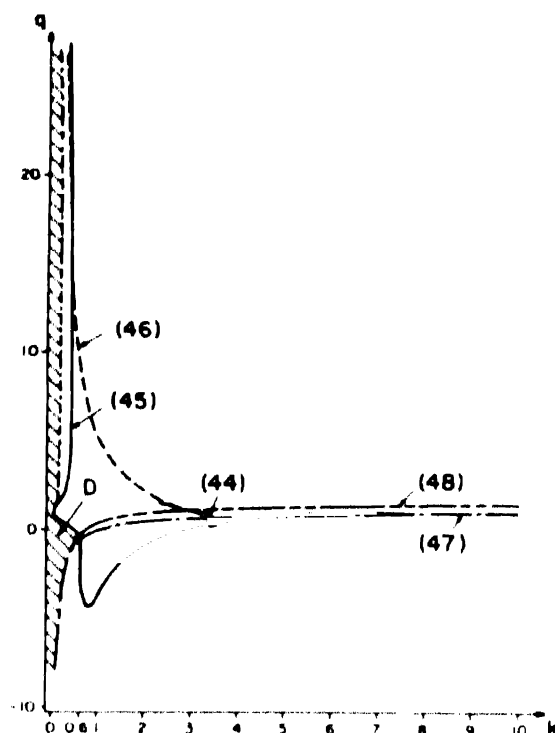


FIG. 3. The complement of the zero set \bar{V} for Example 4D and the connected components of $\mathbb{R}^+ \times \mathbb{R} \setminus L$.

system is absolutely robust stable, according to the Popov criterion, for any nonlinearity ϕ in the sector $S(k)$, and any $r \in [0.001, 1/2]$. Following Fig. 3, we obtain that, the maximal sector in

which our system is absolutely robust stable is $S(0.6)$.

5. CONCLUSIONS

The generalized zero set introduced in Walach and Zeheb (1982) is extended here to continuously differentiable scalar and vector-valued functions, which depends on several real variables and complex parameters. A new method for locating this set is established here.

A design problem for a continuous nonlinear Lurie type system with a linear part under uncertainty conditions is considered. The complete feasible set of sectors of nonlinearities, for which robust absolute stability is ensured, according to the Popov criterion, is found. The generalized zero set method, proposed here, is applied easily in the solution of this problem.

Acknowledgement—The author is greatly indebted to Professor Uri Srebro for carefully reviewing this paper and for his very valuable and constructive comments and suggestions. The author wants to thank the referee for bringing Siljak's work to my attention.

REFERENCES

- Fruchter, G. (1988) Zero sets location and robust stabilization of control systems. Ph.D. dissertation, Technion-Israel Institute of Technology, Haifa, Israel.
- Fruchter, G., U. Srebro and E. Zeheb (1987a) On several variable zero sets and application to MIMO robust feedback stabilization. *IEEE Trans. Circ. Syst.*, **34**, 1208–1220.
- Fruchter, G., U. Srebro and E. Zeheb (1987b) An analytic method for design of uncertain discrete nonlinear control systems. *Proc. 15th Conf. IEEE, Israel*.
- Fruchter, G., U. Srebro and E. Zeheb (1991a) Conditions on the boundary of the zero set and application to stabilization of systems with uncertainty. *J. Math. Anal. Applic.*, **160** (to appear).
- Fruchter, G., U. Srebro and E. Zeheb (1991b) On possibilities of utilizing various conditions to determine a zero set. *J. Math. Anal. Applic.*, **160** (to appear).
- Jury, E. I. and B. W. Lee (1964) On the stability of a certain class of nonlinear sampled-data systems. *IEEE Trans. Aut. Control*, **9**, 51–61.
- Lurie, A. I. (1954) *Some Nonlinear Problems in the Theory of Automatic Control*. Her Majesty's Stationery Office, London.
- Lurie, A. I. and V. N. Postnikov (1944) On the theory of stability of control systems. *Prikl. Mat. Mech.*, **8**, 246–248 (in Russian).
- Popov, V. M. (1961) Absolute stability of nonlinear systems of automatic control. Translated from *Avtomatika i Telemekhanika*, **22**, 961–979.
- Siljak, D. (1969a) Parameter analysis of absolute stability. *Avtomatika*, **5**, 385–387.
- Siljak, D. (1969b) *Nonlinear Systems: The Parameter Analysis and Design*. Wiley, New York.
- Siljak, D. (1989) Polytopes of nonnegative polynomials. *ACC Conf. Pittsburg, PA*, 193–199.
- Walach, E. and E. Zeheb (1982) Generalized zero sets of multiparameter polynomials and feedback stabilization. *IEEE Trans. Circ. Syst.*, **29**, 15–23.

Brief Paper

An Optimal Gas Supply for a Power Plant Using a Mixed Integer Programming Model*

K. AKIMOTO,[†] N. SANNOMIYA,[‡] Y. NISHIKAWA[§] and T. TSUDA^{||}

Key Words—Optimal planning, mixed-integer programming, computer control, production control, energy saving, steel industry.

Abstract—At Kawasaki Steel Mizushima Works, a new energy control system has been established in order to deal with energy and utilities in the steelworks. The system consists of the works' central computer, online computer, process computer and digital instrumentation system. The software system is divided into the planning system, execution system and evaluation system. As the representative topic of the execution system, an optimal gas supply amount for the joint electric power plant is determined. An optimal guidance for the gas supply amount is given to operators by solving a mixed-integer program by a decomposition method in process computer online realtime. This new energy control system has brought a satisfactory energy saving effect.

1. Introduction

IN ENERGY-CONSUMING processes, such as the steelmaking process, the problem of utilizing energy efficiently should not be considered as that of saving energy for each plant. In these processes, various kinds of primary energy are consumed and, at the same time, are converted into by-product energy. A part of necessary energy is supplied from this by-product energy. We then have a complicated interrelationship between generation and consumption of energy. Consequently, the energy saving problem must be considered from the viewpoint of total system management. To this end, computer control systems have been used so far and optimization techniques have been applied for solving this problem.

Kawasaki Steel's Mizushima Works introduced a process computer into its Energy Center about ten years ago. We have, however, experienced such events as severe shortage of petroleum during the oil crisis, greatly changing the situation of the Energy Center. To cope with this situation, Mizushima Works has completely modernized the old computer system and has established a general system which operates in combination with the production control system covering the

entire Works. The new system has been in operation since January 1988.

Various operation control systems were developed in the construction of the new system. This paper describes one of the representative examples, the optimal gas supply system for the power plant, which uses mathematical programming.

As an optimization technique, linear programming has been widely used so far, because we have to deal with a large-scale system. In many cases, a system to be considered has been expressed in a linear programming model (Nishiyama *et al.* 1984, Akagi *et al.* 1986, Ueno *et al.* 1986, Hara *et al.* 1988). On the other hand, a mixed-integer programming model is formulated in order to optimize the combination of various performances of the equipments and to deal with complicated requirements for actual plant operation.

This paper discusses the optimization of the energy supply/demand balance, particularly as it is applied to the by-product gas produced by various processes at the Works and supplied to the joint electric power plant. The amount of surplus gas produced in accordance with the production plan within the Works fluctuates with time. For the joint electric power plant, however, a stable supply of high-caloric gas from the steelworks is desirable, and frequent changing of the fuel for boilers should be avoided so that gas is stored in the gas holder. In consideration of these conditions, it is important to determine an optimum gas supply series, in order to inform the joint electric power plant in advance of what supply of surplus gas will be available, up to a certain point of time.

This problem can be formulated in a mixed integer program problem with a staircase structure. The solution of the problem is obtained by the decomposition method developed by Sannomiya and Okamoto (1985), in which the nested decomposition method in linear program proposed by Grassey (1973) and Ho and Manne (1974) is extended to the mixed-integer program. By repeatedly solving small-scale problems decomposed for each period of notification, it is possible to obtain a feasible suboptimal solution. The validity of the mathematical models and performance of the algorithm are investigated through computation using a large computer for several examples given on the basis of actual operational data. The capability of putting the gas supply plan online by means of the process computer at the site is also studied.

2. Configuration of computer system

Figure 1 shows the hardware configuration of the new system. The hardware system consists of four levels, the central computer (C/C) at the Works, online computer (O/C), process computer (P/C) and the devices included in the digital instrumentation system. This overall system handles such processes as fuel gas, electric power, water, steam and environmental control.

The software of the computer system can be divided into three subsystems, the planning system, execution system and evaluation system. The planning system is installed in the C/C and calculates the supply/demand balance of energy for

* Received 11 February 1990; received in final form 22 September 1990. The original version of this paper was presented at the 6th IFAC Symposium on Mining, Mineral and Metal Processing which was held in Buenos Aires, Argentina during September, 1989. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Editor A. P. Sage.

[†] Kawasaki Steel Corporation Mizushima Works, Kawasaki-Dohri 1, Mizushima, Kurashiki, Okayama 712, Japan. Author to whom all correspondence should be addressed.

[‡] Kyoto Institute of Technology, Kyoto, Japan.

[§] Department of Electrical Engineering, Kyoto University, Kyoto, Japan.

^{||} FUJIFACOM Corporation, Hino, Tokyo, Japan.

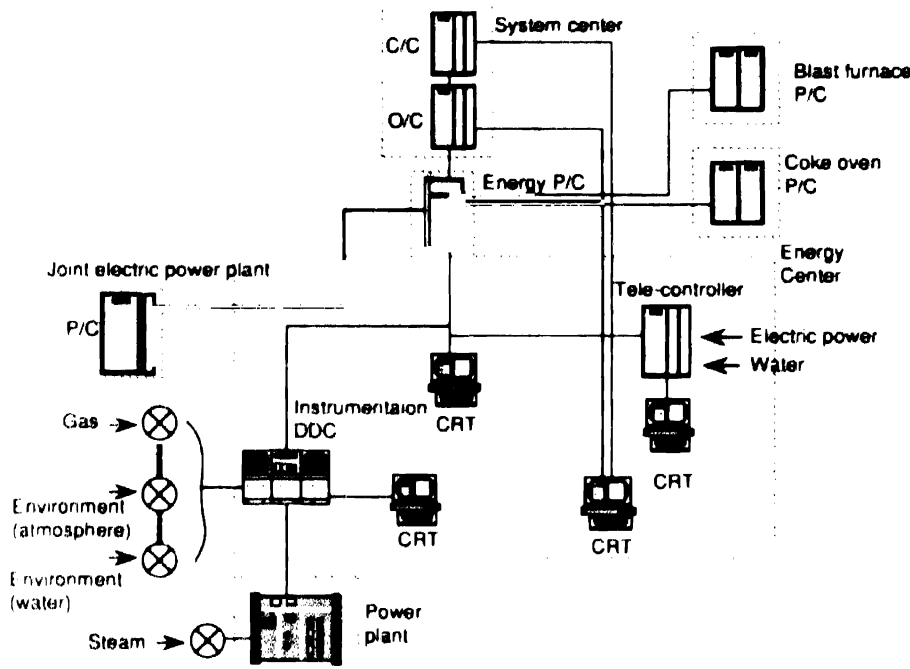


FIG. 1 Hardware configuration of the system

each plant on the basis of information from the production control system of the Works. If there is any difference between demand and supply, the planning system corrects the plan so that the energy balance is maintained. The methodology developed in this paper is utilized in the process computer system as a part of the software for the execution system.

3. Gas supply system

A steelworks generates three types of by-product gas—blast furnace gas (B gas), coke oven gas (C gas) and mixed gas (M gas) such as converter gas. These gases are used within the Works, and the residue is stored in the gas holder or supplied to the joint electric power plant. In using the gas supplied from the Works, it is desirable, from the viewpoint of the joint electric power plant, that the gas supply does not fluctuate with time and that the amount of gas supply should be known well in advance (normally 8 hours ahead). For this reason, the steelworks must inform the joint electric power plant of the future gas supply in advance ("gas notification").

Figure 2 shows the gas supply system discussed here. The left portion of the broken line in the figure is operated on the basis of the Work's production plan, so the optimal gas operation plan to the right of the broken line will be considered here. Gas generation and consumption within the steelworks fluctuates widely over very short times. The balance is maintained on basis of the amount held in the gas holder and the M gas production. M gas is mixed with B gas immediately before use by the joint electric power plant. As a result, the thermal values (in calories) of the B gas increases, so the mixing of M gas is called the calorie-increasing operation.

As shown in Fig. 2, the joint electric power plant has five boilers. Boilers No. 1 and 2 can use calorie-increased B gas and heavy oil as fuel, while boilers No. 3, 4 and 5 can also use C gas. In order to stabilize combustion in the boiler, it is necessary to combine the heavy oil or C gas with calorie-increased B gas. If, therefore, C gas can be supplied at a stationary rate to boilers No. 3, 4 and 5, it is possible to reduce the consumption of heavy oil. Operation of the boiler with gas only, without using heavy oil, is called combustion-of-gas-only. Combustion-of-gas-only reduces the

consumption of heavy oil fuel, leading to a reduction in fuel costs.

The entire period of the plan is divided into T periods corresponding to the gas notification time. Then we determine the series of gas supply notification amounts so as to maximize T period profit when the scheduled gas production and gas consumption in each period at the steelworks are given.

4. Mathematical formulation

Under the assumption that the gas flow rate in any given period is constant, we define the variables in the period $t(1 \leq t \leq T)$ as follows.

$F_{BR}(t)$, $F_{CR}(t)$, $F_{MR}(t)$: Residual gas flow rate in the Works (Nm^3/h) for B gas, C gas and M gas, respectively. These values are known.

$V_B(t)$, $V_C(t)$, $V_M(t)$: Gas volume (Nm^3) held at the end of period t in the gas holder for B gas, C gas and M gas, respectively.

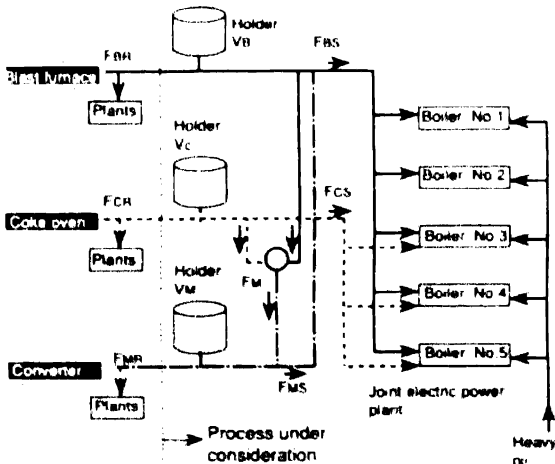


FIG. 2. Gas supply system

$F_M(t)$: M gas flow rate (Nm^3/h) produced from B gas and C gas.

$F_{BS}(t)$, $F_{CS}(t)$, $F_{MS}(t)$: Gas flow rate (Nm^3/h) supplied to joint electric power plant for B gas, C gas and M gas, respectively.

Using these variables, the constraints and objective function of this problem are formulated as follows

(a) *Constraints for gas holder.* The amount of gas held in each holder is given by

$$\left. \begin{aligned} V_B(t) &= V_B(t-1) + \Delta T[F_{BS}(t) - F_{BS}(t) - \alpha F_M(t)] \\ V_C(t) &= V_C(t-1) + \Delta T[F_{CS}(t) - F_{CS}(t) - (1-\alpha)F_M(t)] \\ V_M(t) &= V_M(t-1) + \Delta T[F_{MS}(t) - F_{MS}(t) + F_M(t)] \end{aligned} \right\} \quad (1)$$

where $\Delta T(h)$ is the time interval of period t and α is the mixing ratio of B gas and C gas. The amounts of gas held at the beginning period and the final period of the plan are given by

$$V_B(0) = V_B^0, \quad V_C(0) = V_C^0, \quad V_M(0) = V_M^0 \quad (2)$$

$$V_B(T) = V_B^T, \quad V_C(T) = V_C^T, \quad V_M(T) = V_M^T \quad (3)$$

For the amount held in the holder, there are constraints, i.e. upper and lower limits, which cannot be exceeded. For the sake of convenience, however, the amount (discharge) exceeding the upper limit for each gas is defined as $V_B^+(t)$, $V_C^+(t)$, $V_M^+(t)$ and the amount (gas shortage) lower than the lower limit for each gas as $V_B^-(t)$, $V_C^-(t)$, $V_M^-(t)$. Planning is done so that these values equal zero. In this way, the following constraints for the upper and lower limits are obtained

$$\left. \begin{aligned} V_B - V_B(t) &\leq V_B(t) \leq V_B + V_B^+(t) \\ V_C - V_C(t) &\leq V_C(t) \leq V_C + V_C^+(t) \\ V_M - V_M(t) &\leq V_M(t) \leq V_M + V_M^+(t) \end{aligned} \right\} \quad (4)$$

where V_B , V_C , V_M are the lower limit volumes for the respective gas holders and V_B , V_C , V_M are the upper limit volumes

The amount in the holder is allowed to fluctuate within the range indicated by (4). However, with reference to actual operational results, it is found out that operators keep the gas volume in each holder as constant as possible. Those results should be adopted to reflect operator experience. Thus the normal value of the amount held in each holder is defined as V_B^N , V_C^N , and V_M^N . Using the variation from the normal value, each amount in the holder is expressed as follows

$$\left. \begin{aligned} V_B(t) &= V_B^N + \Delta V_B^+(t) - \Delta V_B^-(t) \\ V_C(t) &= V_C^N + \Delta V_C^+(t) - \Delta V_C^-(t) \\ V_M(t) &= V_M^N + \Delta V_M^+(t) - \Delta V_M^-(t) \end{aligned} \right\} \quad (5)$$

where

$\Delta V_B^+(t)$, $\Delta V_C^+(t)$, $\Delta V_M^+(t) \geq 0$: Increase (Nm^3) from normal value of each gas

$\Delta V_B^-(t)$, $\Delta V_C^-(t)$, $\Delta V_M^-(t) \leq 0$: Decrease (Nm^3) from normal value of each gas

(b) *Constraints for gas flow rate and quantity of heat.* Due to the boiler operation constraints at the joint electric power plant, the following constraints hold for the supply amounts of B gas and C gas.

$$\left. \begin{aligned} F_{BS} &\leq F_{BS}(t) \leq \bar{F}_{BS} \\ F_{CS} &\leq F_{CS}(t) \leq \bar{F}_{CS} \end{aligned} \right\} \quad (6)$$

where F_{BS} and F_{CS} are lower limit values (Nm^3/h) for the supply amount of each gas, and \bar{F}_{BS} and \bar{F}_{CS} the upper limit values (Nm^3/h)

The constraints for the supply amount of M gas, on the other hand, is based on the caloric value of the calorie-increased B gas. The following relation holds from the thermal balance before and after M gas is mixed with B gas

$$q_{B1}[F_{BS}(t) + F_{MS}(t)] = q_{B1}F_{BS}(T) + q_{M1}F_{MS}(t) \quad (7)$$

where q_B and q_M are calories (kcal/Nm^3) of B gas and M gas respectively, and q_{B1} calories (kcal/Nm^3) of B gas after the calorie-increase operation. For q_{B1} , the following upper and lower limits exist

$$q_{B1} \geq \bar{q}_{B1} \geq \hat{q}_{B1} \quad (8)$$

where \bar{q}_{B1} and \hat{q}_{B1} are lower and upper limit values for q_{B1} . From formulae (7) and (8), the following formulae are obtained

$$\left. \begin{aligned} (q_{B1} - \bar{q}_{B1})F_{BS}(t) + (q_{B1} - \bar{q}_{B1})F_{MS}(t) &\leq 0 \\ (q_{B1} - \hat{q}_{B1})F_{BS}(t) + (q_{B1} - \hat{q}_{B1})F_{MS}(t) &\geq 0 \end{aligned} \right\} \quad (9)$$

(c) *Constraints for combustion of C gas only in boilers of the joint electric power plant*

(i) The minimum amount of C gas required for combustion-of-gas-only per boiler is assumed to be F_{CS}^* . Since the combustion-of-gas-only is possible for three boilers, we have the following relationship between the amount of C gas and the number of boilers $n(t)$ for combustion-of-gas-only

$$\left. \begin{aligned} n(t) &= 0 & \text{if } F_{CS} \leq F_{CS}(t) \leq F_{CS}^* \\ n(t) &= 1 & \text{if } F_{CS}^* \leq F_{CS}(t) \leq 2F_{CS}^* \\ n(t) &= 2 & \text{if } 2F_{CS}^* \leq F_{CS}(t) \leq 3F_{CS}^* \\ n(t) &= 3 & \text{if } 3F_{CS}^* \leq F_{CS}(t) \leq F_{CS} \end{aligned} \right\} \quad (10)$$

To express the relationship (10) in a form of the linear inequalities, 0-1 integer variables $n_1(t)$, $n_2(t)$ and $n_3(t)$ are introduced, and the following formulae are thus obtained

$$\left. \begin{aligned} n(t) &= n_1(t) + n_2(t) + n_3(t) \\ F_{CS}(t) &\geq F_{CS}^* - U[1 - n_1(t)] \\ F_{CS}(t) &\leq F_{CS}^* - \delta + Un_1(t) \\ F_{CS}(t) &\geq 2F_{CS}^* - U[1 - n_2(t)] \\ F_{CS}(t) &\leq 2F_{CS}^* - \delta + Un_2(t) \\ F_{CS}(t) &\geq 3F_{CS}^* - U[1 - n_3(t)] \\ F_{CS}(t) &\leq 3F_{CS}^* - \delta + Un_3(t) \end{aligned} \right\} \quad (11)$$

where U is a sufficiently large constant and δ is a properly small positive constant

(ii) If the supply amount of C gas decreases and the combustion-of-gas-only is no longer possible, heavy oil is used as fuel of the boilers. In consideration of the work load required when switching from C gas to heavy oil, frequent switching of fuel is not desirable. When the combustion-of-gas-only is practised, therefore, it should be continued at least during $2\Delta T$ hours. If such continuation is not possible, a penalty should be imposed. In addition, if the number of boilers for combustion-of-gas-only decreases, the number of the decrease should be one at a time. For a simultaneous changeover of multiple boilers, a relatively heavy penalty is imposed.

In order to express this constant, $\Delta n^+(t)$ and $\Delta n^-(t)$ are introduced as variables for the time variation of $n(t)$. That is,

$$\left. \begin{aligned} n(t) &= n(t-1) + \Delta n^+(t) - \Delta n^-(t) \\ \Delta n^+(t), \Delta n^-(t) &\leq 0 \\ n(0) &= n^0, \quad n(T) = \text{free} \end{aligned} \right\} \quad (12)$$

For the decrease of $n(t)$, $\Delta n^-(t)$ is expressed as follows.

$$\left. \begin{aligned} \Delta n^-(t) &= k_1(t) + k_2(t) + k_3(t) \\ k_1(t) &\leq k_2(t) \leq k_3(t) \\ k_1(t), k_2(t), k_3(t) &= 0 \text{ or } 1. \end{aligned} \right\} \quad (13)$$

Furthermore, in order to avoid that both $\Delta n^+(t)$ and $\Delta n^-(t)$ become positive in (12) we have the following constraint

$$0 \leq \Delta n^+(t) + 3k_1(t) \leq 3 \quad (14)$$

When the combustion-of-gas-only cannot be continued for $2\Delta T$ hours, we have the following penalty.

$$p(t) = \begin{cases} \Delta n^-(t) & \text{if } \Delta n^-(t-1) \neq 0 \text{ or } \Delta n^+(t-1) \neq 0 \\ 0 & \text{if } \Delta n^-(t-1) = \Delta n^+(t-1) = 0 \end{cases} \quad (15)$$

In order to express the relationship (15) in a form of linear

inequalities, 0-1 integer variable $l(t)$ is introduced, and the following formulae are thus obtained

$$\begin{aligned} 0 &\leq p(t) \leq \Delta n^+(t) \\ \Delta n^-(t) - U[1 - l(t-1)] &\leq p(t) \\ p(t) &\leq U l(t-1) \\ 0 &\leq M(t) - [\Delta n^-(t) + \Delta n^+(t)] \leq 3 \cdot \\ l(t) &= 0 \text{ or } 1 \\ l(0) &= l^0, l(T) = \text{free.} \end{aligned} \quad (16)$$

(d) *Objective function.* The objective function consists of the following items.

- (i) Evaluated value of gas supply amount.
- (ii) Profit obtained by reducing heavy oil consumption through combustion-of-gas-only.
- (iii) Penalty imposed when the upper and lower limits of gas volume held in holder are violated due to gas discharge or gas shortage
- (iv) Penalty for suppressing fluctuation of gas volume held in holder.
- (v) Penalty imposed when combustion-of-gas-only cannot be continued
- (vi) Penalty imposed when the fuel of multiple boilers is changed simultaneously from C gas to heavy oil.

By summing up the above items, the objective function to be maximized is as follows.

$$\begin{aligned} z_T = \sum_{t=1}^T & [r_1 \Delta T [q_n F_{nS}(t) + q_c (F_{cS}(t) - F_{cS}^0 n(t)) + q_M F_{MS}(t)] \\ & + r_2 \Delta T [n_1(t) + n_2(t) + n_3(t)] \\ & - [r_3 V_B^+(t) + r_4 V_C^+(t) + r_5 V_M^+(t) + r_6 V_B(t) \\ & + r_7 V_C(t) + r_8 V_M(t)] - r_9 p(t) \\ & - [r_{10} \Delta V_B^+(t) + r_{11} \Delta V_B(t) + r_{12} \Delta V_C^+(t) \\ & + r_{13} \Delta V_C(t) + r_{14} \Delta V_M^+(t) + r_{15} \Delta V_M(t)] \\ & - [r_{16} k_2(t) + r_{17} k_3(t)] \end{aligned} \quad (17)$$

where

- q_c : calorie (kcal/Nm³) of C gas
 r_1 : evaluated value of gas (yen/kcal)
 r_2 : profit (yen/h) obtained through combustion-of-gas-only
 $r_3 - r_6$: loss (yen/Nm³) due to gas discharge or shortage at each holder
 r_7 : loss (yen) imposed when the combustion-of-gas-only cannot be continued
 $r_{10} - r_{15}$: penalty (yen/Nm³) for suppressing the fluctuation of gas amount held
 r_{16}, r_{17} : penalty (yen) imposed when the fuel of two or three boilers are changed over simultaneously.

5. Solution

The optimal gas supply plan can be decided by maximizing the objective function (17) subject to the constraints (1)-(6), (9), (11)-(14) and (16). In this section, we discuss an algorithm used to quickly obtain a feasible suboptimal solution to this problem when the sequence of the residual gas amounts ($F_{nS}(t)$, $F_{cS}(t)$ and $F_{MS}(t)$) is given.

The variables defined in the preceding section are classified into 0-1 variables and non-negative continuous variables. If a constraint of the problem includes a variable relating to the adjacent periods, say period t and period $t-1$, the variable is handled as a continuous variable for convenience sake. For this reason, $l(t)$ included in (16) is taken as a continuous variable, and in its stead, 0-1 variable $m(t)$ is introduced. That is, the fifth equation of (16) is substituted by

$$l(t) = m(t), \quad m(t) = 0 \text{ or } 1$$

Thus we define the new variables as follows.

$$\begin{aligned} y(t) &\triangleq [V_B(t), V_C(t), V_M(t), n(t), l(t)]' \\ u(t) &\triangleq [F_{nS}(t), F_{cS}(t), F_{MS}(t), F_M(t), \\ &V_B^+(t), V_C^+(t), V_M^+(t), V_B(t), V_C(t), V_M(t), \\ &\Delta V_B^+(t), \Delta V_B(t), \Delta V_C^+(t), \Delta V_C(t), \Delta V_M^+(t), \\ &\Delta V_M(t), p(t), \Delta n^-(t)]' \\ v(t) &\triangleq [n_1(t), n_2(t), n_3(t), k_1(t), k_2(t), k_3(t), m(t)]' \end{aligned} \quad (18)$$

where $y(t)$ and $u(t)$ are vectors of continuous variables and $v(t)$ is a vector of integer variables. A prime denotes the transition of a vector. Some components of $y(t)$ and $u(t)$ are integer variables, but these components are obtained as integer values from the constraint of the problem. Therefore, they can be treated as continuous variables.

If the variables are defined by (18), the problem in the preceding section is expressed in the following form

$$P: \min z_T = \sum_{t=1}^T [a'y(t) + b'u(t) + c'v(t)] \quad (19)$$

subject to

$$\begin{aligned} y(t) &= Ay(t-1) + Bu(t) + Cv(t) + u(t) \\ Dy(t-1) + Eu(t) + Gv(t) &= d(t) \\ y(t) &\geq 0, \quad u(t) \geq 0 \end{aligned} \quad (20)$$

$$v(t) \in \Theta \triangleq \{v(t) \mid v_i(t) = 0 \text{ or } 1, i = 1, 2, \dots, 7, t = 1, 2, \dots, T\}$$

$$y(0) = y^0 \quad (21)$$

where A, B, C, D, E, G are matrices of appropriate dimension and $a, b, c, u(t), d(t)$ and y^0 are vectors of appropriate dimension. All of these matrices and vectors are known quantities.

The problem P has a constraint with staircase structure, because the coefficient matrix of the second equations of (20) has the non-zero elements distributed in staircase form.

A decomposition algorithm has been proposed for solving the problem P for the case of $D = 0$ (Sannomiya and Okamoto, 1985). With a slight modification, the algorithm is also available for the case of $D \neq 0$.

The procedure is summarized as follows. In a manner similar to the nested decomposition for linear programs (Grasssey, 1973; Ho and Manne, 1974), the problem is decomposed into a series form of T small-size mixed-integer programs. Then, a multi-level technique is applied to the problem with T levels. At each level, the mixed-integer program acts as a master for the following level and as a subproblem for the preceding level. The solution for each level is obtained by solving the small-size problem. The procedure terminates after satisfying a restricted optimality check for each level. Consequently, the algorithm gives a feasible suboptimal solution. The detailed description of the algorithm is omitted here (see Sannomiya and Okamoto, 1985).

The present algorithm does not always give the optimal solution of P , because the optimality of the solution obtained is checked in a restricted manner. Anyway the result obtained at the termination of the algorithm is adopted as the solution of P . The difference between the optimal objective value z_{opt} and the objective value z_T for the current solution is estimated in the following way. By relaxing the first and the second equation of (20), we construct the Lagrange problem, where the Lagrange multipliers are given by the simplex multipliers obtained by the present algorithm. Then, the duality gap ϵ_{max} for this problem is related to z_T and z_{opt} as

$$0 \leq z_T - z_{opt} \leq \epsilon_{max} \quad (22)$$

The value ϵ_{max} is given by

$$\epsilon_{max} = \sum_{t=1}^T \omega_t = \sum_{t=1}^T [\pi_1(t)' u(t) - \pi_2(t)' d(t)] \quad (23)$$

where

$$\omega_i = \min \{ [b + B^T \pi_1(t) - E^T \pi_2(t)]^T u(t) + [c + C^T \pi_1(t) - G^T \pi_2(t)]^T v(t) \} \quad (24)$$

subject to $u(t) \geq 0, v(t) \in \Theta$

In (23) and (24), $\pi_1(t)$ and $\pi_2(t)$ are the simplex multipliers corresponding to the first and second equation for the modified version of (20), respectively. Usually, the value of the duality gap is denoted by $(\epsilon_{\max}/z_T) \times 100$ (%).

6. Example of calculation results

In order to investigate the validity and the effectiveness of the present method, an off-line numerical calculation has been executed for several illustrative examples prepared corresponding to real operation data. The whole planning period is eight hours ($\Delta T = 2$ h and $T = 4$). The numerical computation has been made by a large-scale computer (FACOM M-382 at a Data Processing Center of Kyoto University). The computation has been also made by a process computer for evaluating practical applicability.

Firstly, in order to evaluate the validity of the mathematical model, we applied the decomposition algorithm to the model for various values of $y(0)$, $V_B(T)$, $V_C(T)$ and $V_M(T)$ which are based on the real operation data. As an example, Fig. 3 shows a comparison of the solution obtained by the algorithm with the real operation data. We had zero duality gap for this example. Thus the optimal solution was obtained. Fig. 3(a) shows the real operation data of the holder levels $V_B(t)$, $V_C(t)$, $V_M(t)$ and the gas flow rate $F_{BS}(t)$, $F_{CS}(t)$, $F_{MS}(t)$. Fig. 3(b) shows the optimal solution obtained by the algorithm. It is observed that the holder levels for the operation data change with time in some degree and that the levels for the optimal solution fluctuate very little.

Furthermore, the time variation of the boiler number $n(t)$ in the combustion-of-gas-only operation is given by {3, 2, 2, 3} for the real operation data, and by {3, 3, 3, 3} for the optimal solution. In addition, the objective value obtained by the algorithm was improved by 0.4% for this example as compared with the real operation data. Similar improvement has been obtained for the other examples.

It is concluded from these results that the present mathematical model describes the essential function of the real system with a sufficient degree of reliability. In the case where the parameters in the objective function are set as $r_i = 0$ ($i = 10, \dots, 15$), we have had a solution such as the holder levels change considerably with time within the ranges between the lower and upper bound. The solution is not desirable according to the operator's experience. Thus, it is concluded that the parameters r_i ($i = 10, \dots, 15$) should not be zero.

Secondly, we have had a test run by using the existing process computer (FACOM S-3500 at Energy Center of Kawasaki Steel Mizushima Works, 1.8 MIPS). Computation time was about 14 seconds under no other load condition and 80–100 seconds under normal operation condition. Such computation time is considered to be satisfactory from the viewpoint of practical application. The system gives us a guidance for the gas supply amount automatically every 2 hours corresponding to the gas notification time. The guidance is monitored by operators.

In the Works, the new energy control system has operated since January 1988. It has the gas supply guidance system as a part of software. Figure 4 shows the trend of the combustion-of-gas-only ratio during sixteen months. The ratio increases up to 100% when C gas is surplus. Therefore, in order to use the value as a measure for effectiveness of operation, the ratio divided by F_{CN} (called C gas rate) is shown in the figure.

It is observed from the figure that the utility of C gas has improved successfully since January 1988. Although not all the effect results from this guidance system, it is evaluated that the introduction of the energy control system including the present algorithm has given us quite an improvement.

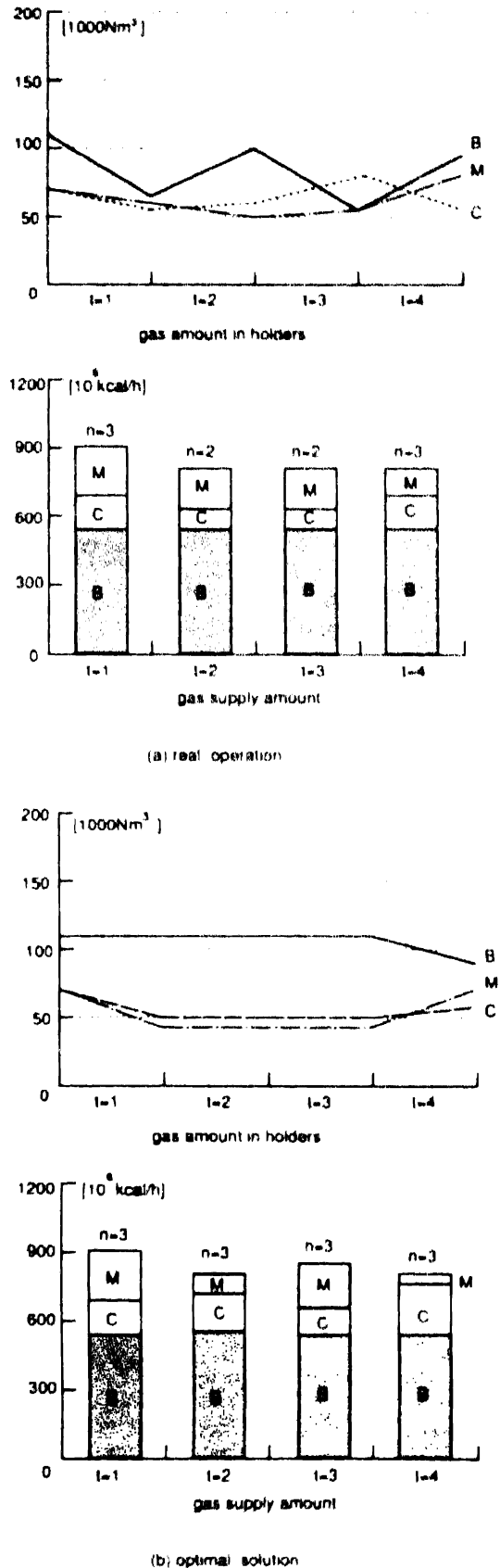


FIG. 3 Comparison of the solution. (a) Real operation; (b) optimal solution

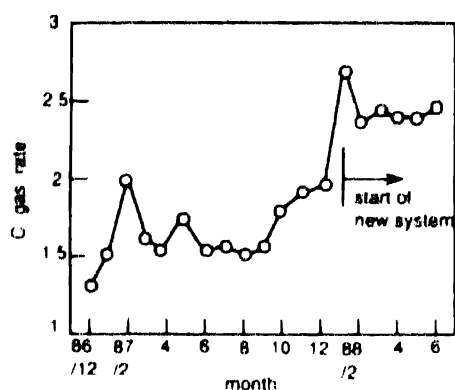


FIG. 4 Trend of the combustion-of-gas-only ratio C gas rate (combustion-of-gas-only ratio/residual C gas amount (10^3Nm^3)). Combustion-of-gas-only ratio = (combustion-of-gas-only time/total time)

7 Conclusion

An optimal gas supply system for a power plant has been discussed in connection with the new energy control system at Kawasaki Steel's Mizushima Works. The problem has been formulated as a mixed-integer program. A suboptimal solution has been obtained by applying a decomposition algorithm. By showing the optimal solution to the operators, sufficient reduction of the heavy oil has been obtained.

References

- Akagi, H., N. Sannomiya, Y. Nishikawa, T. Yashima and T. Tsuda (1986). Production planning and control of an oxygen plant. *Preprints 5th IFAC Symp. on Automation in Mining, Mineral and Metal Processing*, Pergamon, Oxford, pp. 398-403.
- Grassey, C. R. (1973). Nested decomposition and multistage linear programs. *Management Sci.*, **20**, 282-293.
- Hara, H., K. Yamashita and T. Watanabe (1988). Online application of large scale linear programming to pulp and paper mill by mini-computer. *Lecture Notes in Control and Information Sciences* 113. Springer, Berlin, pp. 662-671.
- Ho, J. K. and A. S. Manne (1974). Nested decomposition for dynamic models. *Math. Prog.*, **6**, 121-140.
- Nishiya, T., M. Funabashi and K. Matsumoto (1984). A basic factorization method for multi-state linear programming problems with an application to optimal operation of an energy plant (in Japanese). *Trans. Soc. Instrument and Control Engineers*, **20**, 403-410.
- Sannomiya, N., and K. Okamoto (1985). A method for decomposing mixed-integer linear programs with staircase structure. *Int. J. Syst. Sci.*, **16**, 99-111.
- Ueno, N., Y. Nakagawa, K. Tanimizu and H. Fujisawa (1986). Management and control system for raw materials operation in iron steel industries. *Preprints 5th IFAC Symp. on Automation in Mining, Mineral and Metal Processing*, Pergamon, Oxford, pp. 404-409.

Brief Paper

Robust Measurement Selection*

JAY H. LEE† and MANFRED MORARI‡

Key Words—Robust control, sensors, estimation, observers, disturbance rejection, distillation columns.

Abstract—A measurement selection method is developed in the context of Structured Singular Value (SSV) theory. The method is based on robust performance norm-bounds on transfer functions with direct implications on the sensitivity and robustness of the closed-loop system. Using a high purity distillation column as an example, it is demonstrated that the new measurement selection method can be successfully applied to a realistic problem to yield a measurement set and a controller with robust performance.

1. Introduction

SECONDARY MEASUREMENTS are often an essential aspect of process control. The need for such measurements may stem from one of several factors: first, primary variables may be unmeasurable for either technical or economic reasons; second, secondary measurements may simply improve the system's achievable closed-loop performance. For example, in distillation composition control, temperature sensors often replace expensive, unreliable composition analyzers that introduce significant time delays.

While there is a wealth of both theoretical and practical evidence that points to the importance of correct measurement choice for feedback control system design (leading to desired performance), there is an apparent lack of systematic measurement selection criteria. The importance of measurement selection has long been recognized. A number of different approaches have been proposed for the temperature sensor placement in distillation columns and packed-bed reactors (Bequette and Edgar, 1986; Kumar and Seinfeld, 1978a, b; Moore *et al.*, 1987) as well as in more general contexts (Joseph and Brosilow, 1978; Morari and Stephanopoulos, 1980). Bequette and Edgar (1986) proposed a set of criteria for the temperature sensor placement in distillation columns, based on a compromise between "inferential error" and the measurements' sensitivity to the manipulated variables. Moore *et al.* (1987) proposed a set of empirical rules for the same problem using the singular value decompositions of steady-state gain matrices. Weber and Brosilow (1972) and Joseph and Brosilow (1978) examined the problem of measurement selection in the context of inferential control; their work was extended by Morari and Stephanopoulos (1980) to incorporate system dynamics, based on the Kalman filter. A number of people, notably Kumar and Seinfeld (1978a, b), studied the problem in the stochastic framework and proposed measurement selection criteria minimizing appropriate scalar measures of the covariance matrix of the state-estimation error.

Unfortunately, very few of these proposed criteria have

been tested in the laboratory and far fewer have been applied successfully in industry. The apparent failure of extant selection methods in terms of their general and practical applicability can be attributed to the following facts:

- The methods were restricted to specific types of control schemes or processes.
- The issues of practical importance (such as model uncertainty, system dynamics, unmeasured disturbances/noise, restrictions on the controller structure) are not incorporated.

In this paper, we propose a new measurement selection method that addresses all of the aforementioned practical issues in the context of Structured Singular Value (SSV) theory (Doyle, 1982; Doyle *et al.*, 1982). The proposed selection method is designed to be both numerically simple and for which the subsequent design of robustly performing controllers is straightforward. Skogestad and Morari (1988) showed a means for transforming a necessary and sufficient condition for robust performance in the μ -terminology into a sufficient condition in terms of norm-bounds (maximum singular value) on specific transfer function matrices. We also base our criteria upon norm-bounds on particular transfer function matrices (e.g. sensitivity, complementary sensitivity) with direct implications on the sensitivity and robustness of the closed-loop system.

The method is evidently very general since the SSV formulation allows us to incorporate all the relevant robustness issues as well as all desirable performance features (e.g. user-specified disturbance and performance weights). Furthermore, the subsequent design of the robustly performing controllers is simple since the obtained norm-bounds can be used directly for the controller design. The main drawback of the method is that the derived bounds can be conservative. However, we will show that the conservativeness of the method can be reduced significantly by restricting the controllers to a specific form.

We apply the method to an important chemical process: a high purity distillation column. The placement of temperature sensors in high purity distillation columns is known to be of crucial importance for satisfactory control of the product compositions. The example will demonstrate that the new proposed selection method can be successfully applied to a realistic problem to locate a measurement set for which a robustly performing controller can be easily designed. Although the method is applied to only one process in this paper, we believe that our main contribution lies in the introduction of a general procedure for measurement selection in the face of practical issues including model uncertainty, unmeasured disturbances, measurement noise, and restrictions on the controller structure.

Throughout this paper, it is assumed that all plants can be described by linear, time-invariant, stable transform function matrices. The assumption of stable plants is only for simplicity and the method is not restricted to stable plants.

2. General framework

2.1 Description of disturbances, performance specifications and uncertainties. Figure 1 shows the general diagram of a closed-loop system with secondary measurements and ∞ -norm-bounded uncertainty blocks. We use an input/output

* Received 29 March 1989; revised 3 January 1990; revised 14 May 1990; received in final form 4 August 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak.

† Department of Chemical Engineering, Ross 230, Auburn University, Auburn, Alabama 36849-5127, U.S.A.

‡ Chemical Engineering 210-41, California Institute of Technology, Pasadena, California 91125, U.S.A. Author to whom all correspondence should be addressed.

description of the system where G_{yx} represents the transfer function from the variable x to the variable y . A brief summary of how performance requirements, and model uncertainty is described in the context of structured singular value theory is given in this section

• Description of performance requirements

Performance is measured by the H_∞ -norm of the closed-loop transfer matrix from d' to c' , where the H_∞ norm of a stable, causal (i.e. analytic in the closed RHP) transfer matrix G is defined as

$$\|G\|_\infty = \sup_{\omega} \sigma(G(j\omega)). \tag{2}$$

The H_∞ -norm measures the worst-possible weighted integral square of the output for a class of norm-bounded inputs

• Description of model uncertainty

As shown in Fig. 1, model uncertainty is described as a set of norm (i.e. maximum singular value) bounded perturbations to the nominal model at each frequency. It is assumed that all perturbed models have the same number of RHP poles (hence, stable in our study) as the nominal model

2.2. Structured singular value and robust performance In this section, we summarize the definitions of robust performance and structured singular value and how we may formulate the measurement selection problem in the structured singular value framework

Definition 1 Robust performance The closed loop system shown in Fig. 1 is said to achieve "robust performance" if the closed loop system is nominally stable (i.e. stable with $\Delta_u = 0$) and

$$\max_{\Delta_u \in \Delta_u} \|F_{c'd'}(K, \Delta_u)\|_\infty < 1 \tag{3}$$

where $F_{c'd'}(K, \Delta_u)$ is the closed-loop transfer function matrix from d' to c'

Robust performance implies that the performance specification is satisfied for any plant within the set defined

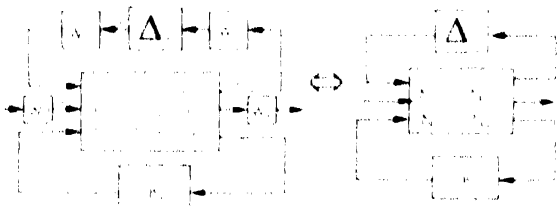


FIG. 1 Formulation of measurement selection problem in the context of structured singular value theory

c = controlled variables

$d = \begin{bmatrix} d \\ n \end{bmatrix}$ = disturbance (d)/noise (n) vector

c' = weighted controlled variables

d' = weighted disturbances/noise

y_j = a (j th) set of measured variables

m = manipulated variables

$\Delta_u = L$, norm-bounded model uncertainty

$$\Delta_u = \left\{ \begin{bmatrix} \Delta_1 & & \\ & \ddots & \\ & & \Delta_r \end{bmatrix} : \Delta_i \in \mathcal{R}^{p_i \times q_i}, \right. \\ \left. \sigma(\Delta_i(j\omega)) \leq 1 \forall \omega \in [0, \infty), i = 1, \dots, r \right\}$$

where \mathcal{R} denotes the set of real-rational transfer function matrices

$$\min_{K \in \mathcal{K}} \min_{\Delta_u \in \Delta_u} \max_{\Delta_u \in \Delta_u} \|F_{c'd'}(K, \Delta_u)\|_\infty < 1$$

$$\min_{K \in \mathcal{K}} \min_{\Delta_u \in \Delta_u} \sup_{\omega} \mu \left[\begin{bmatrix} \Delta_u^* & \\ & \Delta_p \end{bmatrix} \right] (N_{11}(j\omega) + N_{12}K(I - N_{22}K)^{-1}N_{21}(j\omega)) < 1$$

(Note: \mathcal{K} may or may not be constrained to a decentralized structure)

by the uncertainty description. The following theorem by Doyle (1984) shows that a necessary and sufficient condition for robust performance can be formulated in terms of a bound on the SSV (μ) of a particular transfer function matrix (see Appendix for the definition of the function μ).

Theorem 1. Suppose $F_{c'd'}(K, \Delta_u)$ in Fig. 1 is stable when $\Delta_u = 0$. Then

$$\max_{\Delta_u \in \Delta_u} \|F_{c'd'}(K, \Delta_u)\|_\infty < 1 \tag{4}$$

if and only if

$$\sup_{\omega} \mu \left[\begin{bmatrix} \Delta_u^* & \\ & \Delta_p \end{bmatrix} \right] (N_{11}(j\omega) + N_{12}K(I - N_{22}K)^{-1}N_{21}(j\omega)) < 1 \tag{5}$$

where

$$\Delta_u^* = \begin{bmatrix} \Delta_u^* & \Delta_u^* \\ & \Delta_u^* \end{bmatrix} = \begin{bmatrix} \Delta_1 & & \\ & \ddots & \\ & & \Delta_r \end{bmatrix}, \Delta_i \in \mathcal{R}^{p_i \times q_i}$$

$$\Delta_p = \{\Delta : \Delta \in \mathcal{R}^{(\dim(d) + \dim(n)) \times \dim(c)}\} \tag{6}$$

Proof. See Doyle (1984).

2.3 General approaches to measurement selection The following are three possible approaches for measurement selection in the context of SSV theory. The approaches are not mutually exclusive and may be combined in locating the best possible measurement set

Approach 1. Select the measurement set minimizing

$$\min_{K \in \mathcal{K}} \sup_{\omega} \mu \left[\begin{bmatrix} \Delta_u^* & \\ & \Delta_p \end{bmatrix} \right] \times (N_{11}(j\omega) + N_{12}K(I - N_{22}K)^{-1}N_{21}(j\omega)) < 1 \tag{7}$$

where \mathcal{K} is the set of rational transfer function matrices that stabilize the closed-loop system nominally with the particular measurement set under consideration. \mathcal{K} may be restricted to a diagonal/block-diagonal structure if the controller is to be decentralized

Approach 2. Eliminate the measurement sets for which

$$\min_{K \in \mathcal{K}} \sup_{\omega} \mu \left[\begin{bmatrix} \Delta_u^* & \\ & \Delta_p \end{bmatrix} \right] \times (N_{11}(j\omega) + N_{12}K(I - N_{22}K)^{-1}N_{21}(j\omega)) < 1 \tag{8}$$

Approach 3. Find the measurement sets for which

$$\min_{K \in \mathcal{K}} \sup_{\omega} \mu \left[\begin{bmatrix} \Delta_u^* & \\ & \Delta_p \end{bmatrix} \right] \times (N_{11}(j\omega) + N_{12}K(I - N_{22}K)^{-1}N_{21}(j\omega)) < 1 \tag{9}$$

In approach 1, we synthesize K achieving the minimum μ for each measurement candidate, then compare the μ s and select the measurement set and the corresponding controller achieving the lowest μ . Although the approach seems both straightforward and immediate, there are significant theoretical and practical drawbacks to this formulation. The minimization problem expressed through equation 7 is computationally formidable. The algorithm available currently (i.e. μ -synthesis [Doyle, 1984]) is unreliable and requires large CPU-time. In addition, it does not allow restriction of the controller to a decentralized structure. In view of the combinatorial nature of the problem, this approach alone is clearly not a feasible solution to the measurement selection problem for large-scale problems.

In approach 2, necessary conditions for the existence of a controller achieving robust performance are used as screening tools to eliminate the candidates for which no controller exists that achieves robust performance. These screening tools, when "tight" and numerically simple, are of great practical value, since they allow the engineer to reduce the number of measurement candidates dramatically so that more complex criteria may be applied. An example of such screening tools is a condition based on the achievable

H_∞ -norm for nominal performance. If we assume that $\dim(c) = \dim(m)$ and $\dim(s) = \dim(d)$, the following simple screening tool can be used (see Appendix for derivation):

Eliminate the measurement set for which

$$\|[(W_p G_{im})^T (W_p G_{cd} W_d)(G_{sd} W_d)_c], \|_H > 1 \quad (10)$$

where $[\cdot]_*$ denotes the antistable factor, $\|\cdot\|_H$, the Hankel-norm $(\cdot)^*$ an adjoint operator (i.e. $M(s) = M^T(-s)$), and $(\cdot)_c$ and $(\cdot)_{cs}$ inner and coinner factors respectively.

The calculation of the Hankel-norm of a transfer matrix requires solving only two Lyapunov equations. The condition (10) is a necessary and sufficient condition for the existence of K meeting performance requirement in the absence of model uncertainty; hence it is a necessary condition for the existence of K achieving robust performance. For cases where $\dim(m) < \dim(c)$ and/or $\dim(s) < \dim(d)$, an iterative search procedure (e.g. γ -iteration (Doyle, 1984)) is required to calculate the optimal H_∞ -norm for nominal performance. For more of such screening tools, readers are referred to Lee (1991).

In Approach 3, sufficient conditions are used to locate a measurement set for which a controller achieving robust performance exists. There may exist more than one measurement set with this property. In this case, we can either impose more stringent performance specifications, or select a measurement set for which a robustly performing controller can be designed most easily. Development of "light" sufficient conditions useful for measurement selection and controller design is the focus of this paper. In Section 2.4, we will develop a methodology to derive robust performance norm-bounds on particular transfer matrices (that parametrize the controller K). These norm-bounds are sufficient conditions for robust performance and can be used to select a measurement set for which a robustly performing controller can be designed in a straightforward manner.

2.4 Methodology for deriving robust performance norm-bounds In this section, we present briefly the method for deriving robust performance norm-bounds on desired transfer function matrices. We make use of the following theorem by Skogestad and Morari (1988).

Theorem 2. Let $M \in \mathbb{C}^{n \times m}$ be written as

$$M = R_{11} + R_{12}L(I - R_{22}L)^{-1}R_{21} \quad (11)$$

where

$$R_{11} \in \mathbb{C}^{n \times m}, \quad R_{12} \in \mathbb{C}^{n \times p}, \quad R_{21} \in \mathbb{C}^{k \times m}, \quad R_{22} \in \mathbb{C}^{k \times p} \quad \text{and } L \in \mathbb{C}^{p \times k}. \quad (12)$$

Define

$$f(c_l) = \mu \left\| \begin{bmatrix} \Delta & \\ & \Delta_l \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ c_l R_{21} & c_l R_{22} \end{bmatrix} \right\| \quad (13)$$

where

$$\Delta = \left\{ \Delta : \Delta = \begin{bmatrix} \Delta_1 & & \\ & \ddots & \\ & & \Delta_l \end{bmatrix}; \sum_{i=1}^l m_i = m, \sum_{i=1}^l n_i = n, \Delta_i \in \mathbb{C}^{n_i \times m_i} \right\} \quad (14)$$

$$\Delta_l = \left\{ \Delta : \Delta \in \mathbb{C}^{p \times k}, c_l \Delta \in \mathbb{R}^+ \right\}$$

Assume

$$\mu_A(R_{11}) < 1 \quad \text{and} \quad \det(I - R_{22}L) \neq 0 \quad (15)$$

then

$$\mu_A(M) < 1 \quad (16)$$

if

$$\delta(L) < c_l^* \quad (17)$$

where c_l^* solves $f(c_l^*) = 1$.

Proof. See Skogestad and Morari (1988).

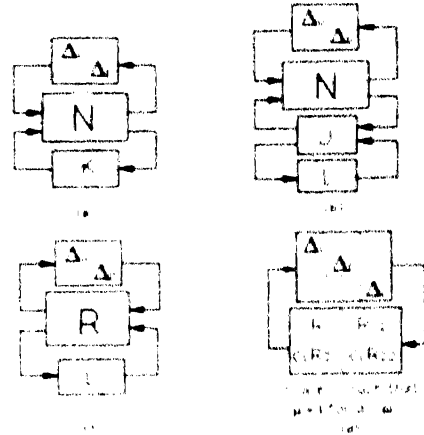


Fig. 2 General method for deriving norm-bounds (a) general robust performance problem, (b) parametrization of K in terms of L , (c) parametrization of M in terms of L , (d) derivation of norm-bounds on L .

The procedure for deriving the norm-bounds is summarized as follows.

Step 1. Figure 2(a) represents the general robust performance problem (see Theorem 1). The first step is to parametrize the controller K in terms of the transfer function matrix L on which the norm-bound is desired, as shown in Fig. 2(b).

Step 2. After combining the matrices N and J into R [Fig. 2(c)], the matrix L is treated as a real scalar parameter c_l times an uncertainty block Δ_l as shown in Fig. 2(d).

Step 3. The norm bound on L , guaranteeing robust performance is calculated by choosing the frequency-dependent scaling factor $c_l^*(\omega)$ such that $f(c_l^*(\omega)) = 1 \forall \omega$. According to Theorem 2, the procedure provides a frequency-by-frequency bound on $\delta(L(j\omega))$ guaranteeing robust performance (assuming K is chosen such that the closed-loop system is nominally stable). We emphasize that $c_l^*(\omega)$ is the *tightest* norm-bound on L in the sense that, for each $c_l(\omega) > c_l^*(\omega)$ for some ω , there exists at least one L such that $\delta(L(j\omega)) = c_l(\omega)$ and $\sup f(c_l(\omega)) > 1$.

Remarks

- $f(c_l)$ is a nondecreasing function of c_l . Thus the scaling factor c_l^* can be easily found through a simple search-procedure (e.g. bisection method).
- There may exist many acts of J, L parametrizing K . The norm-bounds on different L s can be combined over different frequency ranges. For example, suppose that both L_1 and L_2 parametrize K . Then, robust performance is met if, for each ω , $\delta(L_1(j\omega)) < c_{l1}^*(\omega)$ or $\delta(L_2(j\omega)) < c_{l2}^*(\omega)$.
- Sufficiency of the condition (17) arises from the fact that robust performance must be guaranteed for every L satisfying $0 < \delta(L(j\omega)) < c_l^*(\omega)$ (as opposed to a particular L).
- $\mu_A(R_{11}) < 1$ requires that $f(c_l) < 1$ for $c_l = 0$.
- Tightest bounds can be obtained if we restrict L to be a scalar-times-identity matrix. Then, μ can be calculated with respect to $\Delta_l = \{\delta I^{p \times k} : \delta \in \mathbb{C}\}$.

3. Measurement selection criteria based on robust performance norm bounds

Two philosophically different approaches are possible for controlling variables through secondary measurements:

Option 1. Select the secondary measurements with input-output behavior similar to that of the primary variables and control the selected secondary variables.

Option 2. Use the secondary measurements to estimate the primary variables and control these estimates.

Option 1 can be advantageous over option 2 in that the required controller design effort and controller complexity for option 1 may be significantly less than for option 2.

However, option 2 is more general, since there may not exist any secondary measurement set with similar input/output behavior

In this section, we use the robust performance norm-bounds introduced in Section 2.4 to develop measurement selection methods for the above two options. The main question concerns the choice of the transfer matrices on which the robust performance norm-bounds are to be derived.

3.1 Measurement selection criteria for option 1. For this approach, the objective is to select measurements with input-output behavior similar to that of the primary variables so that the controller design can be based on simple techniques such as shaping of the sensitivity and complementary sensitivity functions.

Choice of functions for robust performance norm-bounds. It is logical to make measurement selections based on the robust performance norm-bounds on the sensitivity function $S = (I + G_{im}K)^{-1}$ and the complementary sensitivity function $T = G_{im}K(I + G_{im}K)^{-1}$, since these functions have direct implication on the closed-loop response of the secondary variables. The secondary measurement sets with substantially different input/output behavior from that of the primary variables will yield infeasible bounds, since shaping of these functions is unlikely to result in a controller achieving robust performance.

Figure 3(a) represents a parametrization of the controller K in terms of the complementary sensitivity function T . The parametrization of K in terms of the sensitivity function S can be obtained by using the fact the $S = I - T$. The robust performance norm-bounds on T and S can then be calculated by using the procedure described in Section 2.4. However, there are serious drawbacks in making measurement selection based on these norm-bounds. First, $(G_{im})^{-1}$, a right inverse of G_{im} , may not exist. Second, the stability and causality of T (and hence S) does not necessarily imply the internal stability of the closed-loop system and the causality of the controller K . In fact, in order to have internal stability and controller causality, the transfer function $(G_{im})^{-1}T(s)$ has to be stable and causal. Hence, T is restricted to share the same nonminimum-phase characteristics with G_{im} . This restriction on T may put severe limitations on the achievable $\sigma(S(j\omega))$ for some measurement sets, and the bounds may not provide a meaningful basis for selection.

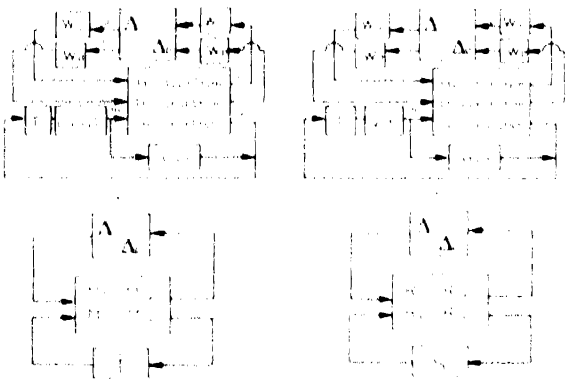


FIG. 3 Parametrization of: (a) controllers in terms of complementary sensitivity function T ; (b) controller in terms of IMC filter \tilde{T} ; (c) closed-loop operator M in terms of \tilde{T} ; (d) closed-loop operator M in terms of $S = I - \tilde{T}$

$$\begin{aligned} M_{11} &= \begin{bmatrix} W_c G_{cu} W_u & W_c G_{cd} W_d \\ W_p G_{pu} W_u & W_p G_{pd} W_d \end{bmatrix} & M_{12} &= \begin{bmatrix} W_c G_{im} Q_{IMC} \\ W_p G_{im} Q_{IMC} \end{bmatrix} \\ M_{21} &= [-G_{su} W_u \quad -G_{sd} W_d] & M_{22} &= 0 \\ R_{11} &= M_{11} + M_{12} M_{21} & R_{12} &= M_{12} \\ R_{21} &= -M_{21} & R_{22} &= 0 \end{aligned}$$

We can overcome these problems by replacing $(G_{im})^{-1}$ in Fig. 3(a) with the IMC (Internal Model Control) controller Q_{IMC} and deriving the norm-bounds on \tilde{T} and $\tilde{S} = I - \tilde{T}$ instead [see Fig. 3(b)]. An explicit formula for H_2 -optimal Q_{IMC} in the case of zero input weighting exists and can be found in Morari and Zafriou (1989). This particular choice of Q_{IMC} corresponds to an approximate stable inverse of $(G_{im})_r$, and hence, \tilde{T} can be viewed as an "approximate" complementary sensitivity function. Indeed, if $(G_{im})_r^{-1}$ exists and is stable, the optimal Q_{IMC} is G_{im}^{-1} yielding $\tilde{T} = T$ and $\tilde{S} = S$. More general H_2/H_∞ optimization techniques (Kwakernaak and Sivan, 1972; Francis, 1987) may be applied to obtain Q_{IMC} if necessary. Because the resulting Q_{IMC} is always stable and causal, the only requirement for internal stability and controller causality is the stability and causality of \tilde{T} . Hence, the robust performance norm-bounds on \tilde{T} and \tilde{S} can be used directly in controller synthesis, and they provide a fair basis for measurement selection. The norm-bounds are easily derived using the method described in Section 2.4, as shown in Fig. 3(b)-(d). The most nonconservative bounds are derived by restricting \tilde{T} and \tilde{S} to be scalar-times-identity transfer matrices.

Steady-state condition for existence of feasible bounds. Although numerically straightforward, it may be cumbersome to derive these norm-bounds for all candidate sets if the number of candidate sets is very large. We can reduce the number of candidate sets which have to be considered significantly by noting that the feasible bounds can be obtained only if the robust performance condition is satisfied at $\omega = 0$ with $S = 0$.

Theorem 3. (Referring to Fig. 5), $c_1^*(0) > 0$ if and only if

$$\mu \left[\begin{bmatrix} \Delta_u^* \\ \Delta_p \end{bmatrix} \right] (R(0)) < 1 \tag{18}$$

where

$$R = \begin{bmatrix} W_c(G_{cu} - G_{im}Q_{IMC}G_{cu})W_u & W_c(G_{cu} - G_{im}Q_{IMC}G_{cd})W_d \\ W_p(G_{pu} - G_{im}Q_{IMC}G_{pu})W_u & W_p(G_{pu} - G_{im}Q_{IMC}G_{pd})W_d \end{bmatrix} \tag{19}$$

Proof. Note that $f(c_1)$ in Theorem 2 is a monotonically non-decreasing function of c_1 . Hence, in order for $c_1^*(0) > 0$, it is necessary and sufficient that the robust performance condition is satisfied at steady state with $S = 0$. From Fig. 3(d), it is evident that this corresponds to the condition (18).

We cannot expect a feasible bound on $\sigma(\tilde{T}(0))$ because $\tilde{T} = 0$ implies open-loop. Since the condition (18) can be checked very easily, it can be used to eliminate efficiently the measurement sets for which the resulting robust norm-bounds on \tilde{T} and \tilde{S} will not be feasible. It is then necessary to derive frequency dependent norm-bounds only on those measurement sets that satisfy the condition (18).

If $Q_{IMC}(0) = (G_{im})_r^{-1}(0)$, $\tilde{S} = 0$ ($\tilde{T} = I$) implies "perfect" steady-state control of the secondary variables s . This is the case if the controller K has integral action on s . Hence, the condition asks whether or not steady-state performance specifications will be met when a controller with integral action (such as diagonal PID) is implemented. It is interesting to note that, if there exists a direct linear relationship between the primary variable vector c and the secondary variable vector s (i.e. $c = Ms$ where M is a constant matrix), then the condition (18) reduces to a much simpler condition

$$\mu \left[\begin{bmatrix} \Delta_u^* \\ \Delta_p \end{bmatrix} \right] [W_c(G_{cu} - G_{im}Q_{IMC}G_{cu})W_u(0)] < 1. \tag{20}$$

Note that the condition (20) does not depend on the performance weight W_p and the disturbance weight W_d . Since the amount of uncertainty (expressed through W_c and W_u) is relatively "small" at steady state, the condition (20) will almost always be satisfied.



FIG. 4. Parametrization of decentralized controller K_d in terms of T_d (a) and S_d (b)

Comparison with existing measurement selection criteria. Next we want to compare our criteria with other available measurement selection criteria. When the uncertainty is ignored ($W_i = 0$, $W_o = 0$) and $Q_{IMC}(0) = (G_{im})^{-1}(0)$, the left-hand side of the inequality (18) becomes $\sigma\{W_p(G_{id} - G_{im}(G_{im})^{-1}G_{id})W_d\}$. This is equivalent to the maximum singular value of the "inferential error" matrix that Bequette and Edgar (1986) suggested to minimize by measurement selection. Since our criteria account for model uncertainty and system dynamics, they are more general and complete. In addition, Bequette and Edgar (1986) stated in their article, that the measurement selection should be based on the compromise between inferential accuracy and sensitivity of measurements to manipulated variables. The sensitivity of measurements to manipulated variables should not impact measurement selection since the amount of control action needed to eliminate a disturbance in the primary variables (which is the ultimate objective) does not depend on the choice of measurements. However, an issue of great importance is the sensitivity of measurements to disturbances. In our criteria, disturbances could include measurement noise so that a compromise between the inferential accuracy and the signal/noise ratio of the measurements is reached.

Extension to decentralized control. The above approach can be easily extended to cases where the controllers are restricted to diagonal/block-diagonal structures ($K_d = \text{diag}\{(K_d)_1, \dots, (K_d)_n\}$). Figure 4(a) represents a parametrization of the decentralized controller K_d in terms of "block-diagonal complementary sensitivity" function $T_d = \text{diag}\{(T_d)_1, \dots, (T_d)_n\}$ where $(T_d)_i = (G_{im})_i(K_d)_i(I + (G_{im})_i(K_d)_i)^{-1}(G_{im})_i$, $1 \leq i \leq n$, represents the i th diagonal block of G_{im} . Using the same argument as before, we choose to derive norm bounds on \tilde{T}_d and $\tilde{S}_d = I - \tilde{T}_d$ instead of T_d and S_d as shown in Fig. 4(b). A condition similar to (18) can be also derived straightforwardly, we omit further details for the sake of brevity.

3.2. Measurement selection criteria for option 2. In this approach, the controller design is based directly on the performance of the primary variables using their estimates from the secondary measurements. Assuming a large enough number of linearly independent measurements (greater or equal to the number of linearly independent disturbances), "perfect" estimation of the primary variables is possible in ideal situations (that is, in the absence of measurement noise, model uncertainty, and nonminimum-phase characteristics in the measurement responses to disturbances). Hence, the objective is to select the measurement set that is least sensitive to these factors preventing "perfect" estimation.

Choice of functions for robust performance norm-bounds. In order to apply the robust norm-bound method, we must seek transfer matrices with direct implications on the performance of the primary variables. Figure 5(a) represents the parametrization of the controller K in terms of the function H . It represents the transfer function from the setpoints r to the primary variables c , and hence we will refer to it as the "complementary sensitivity-like" function. Similarly, we can define the "sensitivity-like" function P as

$I - H$. P represents the transfer function from d to c when the effects of the disturbances on the controlled variables are normalized (i.e. $G_{id} = I$). These functions determine the closed-loop response of the primary variables and have direct relevance for closed-loop robustness as well. Hence, it is logical to base the measurement selection on the robust performance norm-bounds on these functions.

As was the case with sensitivity and complementary sensitivity functions, however, there are serious drawbacks in making the measurement selection based on the norm-bounds on P and H . Firstly, $(G_{id})_i^{-1}$, a left inverse of G_{id} , and $(G_{im})_i^{-1}$, a right inverse of G_{im} , may not exist. Secondly, the stability and causality of H does not necessarily imply the internal stability of the closed loop system and the causality of the controller K . Due to the internal stability and controller causality requirements, the transfer function H is restricted to those transfer matrices yielding stable and causal $(G_{im})_i^{-1}HG_{id}(G_{id})_i^{-1}$. This restriction may put severe limitations on achievable $\delta(P(j\omega))$ for some measurement sets and the bounds may not provide a fair basis for selection.

We can overcome these problems similarly as we did for Option 1. We replace $(G_{im})_i^{-1}$ and $G_{id}(G_{id})_i^{-1}$ in Fig. 5(a) with the IMC controller Q_{IMC} and the estimator E in Fig. 5(b). Then we can use the norm-bounds on H and $P \approx I - H$ instead for measurement selection [see Fig. 3(b)]. The advantage of this is that the only requirement for internal stability and controller causality is the stability and causality of H . Hence, the robust performance norm-bounds on H and P can be used directly in controller synthesis, and they also provide a fair basis for measurement selection. A good

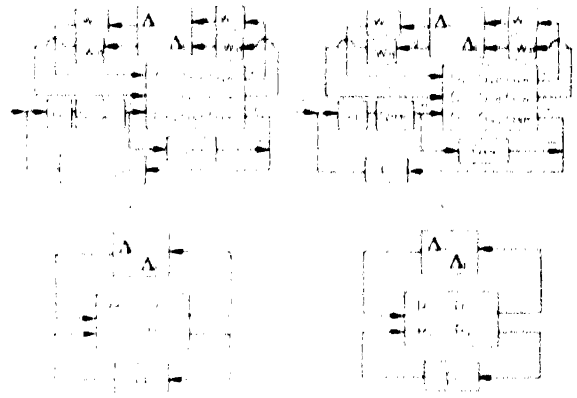


FIG. 5. Parametrization of: (a) controller in terms of "complementary sensitivity-like" function H , (b) controller in terms of IMC filter H , (c) closed-loop operator M in terms of H , (d) closed-loop operator M in terms of $S = I - H$

$$M_{11} = \begin{bmatrix} W_o G_{iu} W_u & W_o G_{id} W_d \\ W_p G_{iu} W_u & W_p G_{id} W_d \end{bmatrix} \quad M_{12} = \begin{bmatrix} W_o G_{im} Q_{IMC} \\ W_p G_{im} Q_{IMC} \end{bmatrix}$$

$$M_{21} = [-EG_{iu} W_u \quad -EG_{id} W_d] \quad M_{22} = 0$$

$$R_{11} = M_{11} + M_{12}M_{21} \quad R_{12} = M_{12}$$

$$R_{21} = -M_{21} \quad R_{22} = 0$$

choice for Q_{IMC} is the H_2 -optimal Q_{IMC} with no input-weighting, although other more general H_2/H_∞ optimization technique can be employed. It corresponds to an approximate, stable inverse of G_m . A good choice for E is $G_d(G_{id})^{-1}$ where $(G_{id})_{ii}$ represents the coouter factor of the coinner-coouter factorization of G_d (such that $(G_{id})_{ii}(0) = I$) (Francis, 1987). However, more generally, E can be chosen as the Kalman filter (or modified Kalman filter). For details on the design of E , the readers are referred to Morari and Stephanopoulos (1980). The optimal choice (in the absence of input-weighting and measurement noise) for Q_{IMC} and E will be $(G_m)_r^{-1}$ and $G_d(G_{id})_i^{-1}$ respectively, if they are both stable and causal.

Steady-state condition for existence of feasible bounds. As in Section 3.1, we can obtain a necessary condition for the existence of a feasible bound on P at $\omega = 0$.

Theorem 4 (Referring to Fig. 5) $\epsilon_P(0) > 0$ if and only if

$$\mu \begin{bmatrix} \Delta_u^* \\ \Delta_p \end{bmatrix} (\hat{R}(0)) < 1 \tag{21}$$

where

$$\hat{R} = \begin{bmatrix} W_u(G_{iu} - G_{im}Q_{IMC}EG_{iu})W_u \\ W_p(G_{iu} - G_{im}Q_{IMC}EG_{iu})W_u \\ W_u(G_{id} - G_{im}Q_{IMC}EG_{id})W_d \\ W_p(G_{id} - G_{im}Q_{IMC}EG_{id})W_d \end{bmatrix} \tag{22}$$

Proof. Straightforward from the proof of Theorem 3.

We cannot expect a feasible bound on $\hat{o}(H(0))$ because $H \approx 0$ implies open-loop. Using the condition (21), we can dramatically reduce the number of the measurement sets for which the norm-bounds are to be derived. If $Q_{IMC}(0) \approx (G_m)_r^{-1}(0)$ and $E(0) \approx G_d(G_{id})_i^{-1}(0)$, $P(0) \approx O(H(0) \approx I)$ implies "perfect" steady state control of the primary variables ϵ in the absence of model uncertainty and measurement noise. Hence, the left-hand side of the inequality (21) is the measure of the sensitivity to uncertainty and measurement noise.

Comparison with existing measurement selection criteria. We would like to draw a comparison between our criteria and the "condition number criterion" proposed by Weber and Brosilow (1972). The minimization of the left-hand side of the inequality (21) with $Q_{IMC}(0) \approx (G_m)_r^{-1}(0)$ and $E(0) \approx G_d(G_{id})_i^{-1}(0)$ is equivalent to the condition number criterion if we assume no measurement noise and a specific type of uncertainty, namely multiplicative "full" output uncertainty on G_d [see Lee (1991) for detail]. Our criteria are more complete since they incorporate a more general description of uncertainty, measurement noise and system dynamics.

4. Numerical example: High-purity distillation

As an example, we study the high-purity distillation column shown in Fig. 6. The column and the model are described in detail in Appendix A of Morari and Zafritou (1989). The control problem of the column is presented in Fig. 7.

Problem description

• Disturbances/noise

Most common disturbances are those in the feed, it often changes according to the conditions in another plant unit such as a reactor. Measurement error (noise) is often another important factor. We will study the effect of one physically motivated measurement error: uncompensated pressure variation. The following set of disturbances/noise is considered:

- Feed flowrate (F)
- Feed composition (z_F)
- Uncompensated pressure variation (P)

• Measured variables

Measurements are usually not limited to a specific number

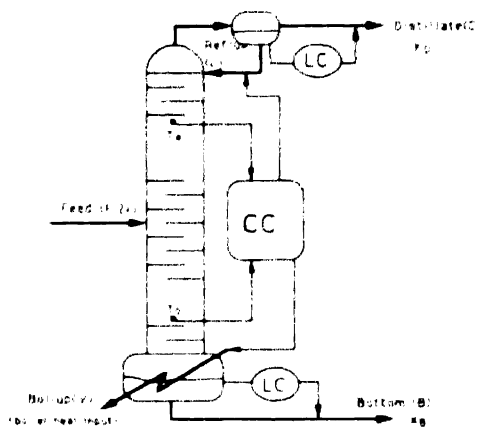


FIG. 6. High-purity distillation column

although it is common to use two tray temperatures for two-point composition control. In this example, for the sake of simplicity, we restrict ourselves to two tray temperatures (T_u and T_b). In addition, for brevity of presentation, we consider only the placements symmetric with respect to the feedtray (such as tray #1/tray #41, tray #2/tray #40, and so on). This is logical since the column is symmetric with respect to the feedtray.

• Uncertainty

We limit ourselves to uncertainty in the manipulated variables. They have been shown to be the dominant uncertainty for high-purity distillation columns (Skogestad *et al.*, 1988). We choose the same uncertainty weight W_f that Skogestad *et al.* (1988) used in their study

$$W_f = 0.2 \frac{5s + 1}{0.5s + 1} I \tag{23}$$

• Performance and disturbance weight

The performance weight W_p and the disturbance weight W_d are chosen as follows.

$$W_p = 0.38 \frac{100s + 1}{10s + 0.01} I \tag{24}$$

$$W_d = [I \quad W_n]$$

where

$$W_n = 0.04 \frac{5s + 1}{0.125s + 1} \begin{bmatrix} \left(\frac{dT}{dP} \right)_{T=T_u} \\ \left(\frac{dT}{dP} \right)_{T=T_b} \end{bmatrix} \tag{25}$$

As usual, much tighter specifications are imposed in the low frequency region in order to ensure good steady-state response.

• Choice of functions for robust-performance norm-bounds

We take the more general approach of Option 2 and choose the sensitivity-like and complementary sensitivity-like functions described in Section 3.2.

Steady-state performance. We apply the steady-state condition (21) to reduce the number of measurements to consider. The plot of the left-hand side of the inequality (21) vs measurement sets is shown in Fig. 8(a). It represents the measure of the "worst-possible" performance when the controller is designed yielding no steady-state offsets in

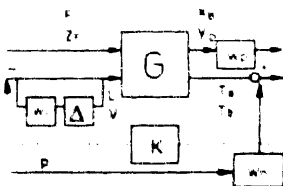


FIG. 7. Control problems in the distillation column.

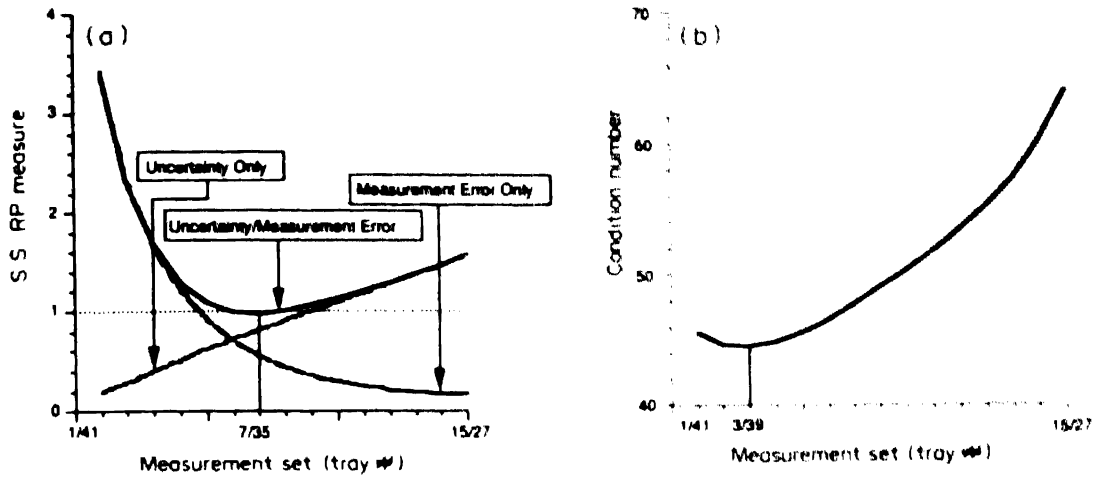


FIG. 8. Robust performance measure at the steady-state with controllers yielding no steady-state offsets in compositions nominally ($P = 0$) (a) Steady-state robust performance measure, (b) condition number of $G_{ss}(0)$

compositions nominally (in the absence of uncertainty and measurement error). The measurement set of T_1 and T_{15} shows the best steady-state performance. In fact, it is the only measurement set that satisfies the condition (21). This result can be interpreted physically. The temperatures measured close to the reboiler and the condenser have poor signal/noise ratio because the gains from disturbances to these measurements are "small". On the other hand, the measurements far away from the reboiler and the condenser become less direct. Hence, placement of the temperature sensors involves a compromise between these two factors. This is apparent from the plots shown in Fig. 8(a) that represent the values for the left-hand side of the inequality (21) when measurement error (uncompensated pressure variation)/model uncertainty are neglected. The measurement set T_7/T_{15} is apparently the best compromise between the signal/noise ratio and the sensitivity to model uncertainty. Note that neglecting either the model uncertainty or the measurement error would have resulted in a wrong choice of measurement. Figure 8(b) represents the

condition numbers of the steady-state gain matrices from the disturbances to the measurements ($G_{ss}(0)$). Note also that the condition number (Bronlow's criterion) does not reflect the measurements' sensitivity to uncertainty correctly in this particular problem.

Robust performance norm-bounds on P and H . Since the only measurement set satisfying the steady-state condition (21) is T_7/T_{15} , all that is left to do is to check that, for this particular measurement set, we can design H such that robust performance is achieved. This can be easily done by deriving the robust performance norm-bounds on $|g(j\omega)|$ and $|1 - g(j\omega)|$. The robust performance bounds on P and H for the measurement set T_7/T_{15} are shown in Fig. 9(a). The bounds are "feasible" since the following transfer function meets at least one of the bounds at every frequency as we can see from Fig. 9(a).

$$H(s) = g(s)I = \frac{107.5s + 1}{(100s + 1)(7.5s + 1)(2s + 1)}I \quad (26)$$

The μ -plot for robust performance [Fig. 9(b)] shows that

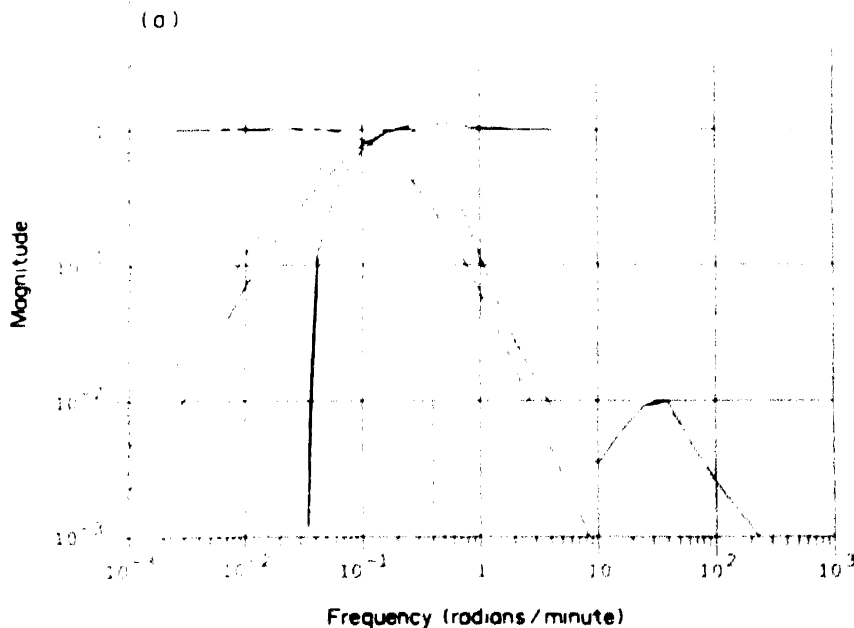


FIG. 9(a)

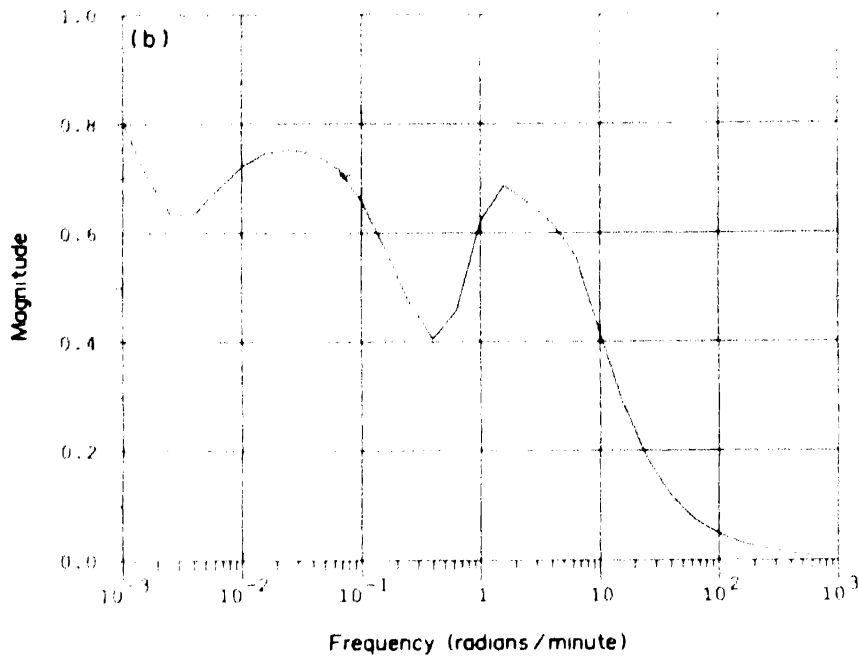


FIG. 9(b)

FIG. 9. Meeting the robust performance norm-bounds on $\delta(\hat{P}(j\omega))$ and $\delta(\hat{H}(j\omega))$ with $\hat{H} = g(s)I$ for the measurement set T_7/T_{35} . (a) Meeting the norm-bounds; (b) SSV for robust performance.

--- bound on $\delta(\hat{P}(j\omega))$ — bound on $\delta(\hat{H}(j\omega))$
... $|1 - g(j\omega)|$ -.- $|g(j\omega)|$

robust performance is achieved for the measurement set T_7/T_{35} . Figure 10 shows the simulated responses of x_B and y_D to unit step disturbances in z_F and F and a measurement noise in the form of a pseudo-random binary signal of unit magnitude filtered through W_n . The specific multiplicative input uncertainty (i.e. $W_I\Delta_I$) used for the simulation is $\begin{bmatrix} 0.2 & 0 \\ 0 & -0.2 \end{bmatrix}$. The simulations confirm the physical interpretation given earlier.

Although not shown here, the option 1 approach (selection based on the norm-bounds of the sensitivity and the complementary sensitivity functions) identified the exactly the same measurement set as the best measurement set in this case.

5. Conclusion

This paper introduces a general measurement selection method in the face of practical issues (such as model uncertainty, restrictions on controller structure, and so on), cast in the framework of the SSV theory. It differs from most of the previous work in its theoretical rigor and general and practical applicability. The method is based upon norm-bounds of the transfer function matrices with direct implication on the sensitivity and the robustness of the closed-loop system. Hence, the derived norm-bounds are useful not only for measurement selection but also for subsequent robust controller design. The choice of matrices on which the norm-bounds are based depends on design philosophy. We identified two philosophically distinct approaches to controlling variables through secondary measurements, and showed how the robust performance norm-bounds on certain matrices can be used as the basis for measurement selection. The example of a high-purity distillation column demonstrated that the proposed selection method can be applied successfully to a practical problem in locating a measurement set leading to robust performance. As a final note, we would like to point out that, although we developed this paper in the context of measurement selection, the proposed norm-bound method is applicable to more general control structure selection problems. For example, it can be used to select an appropriate set of actuators among the available candidates.

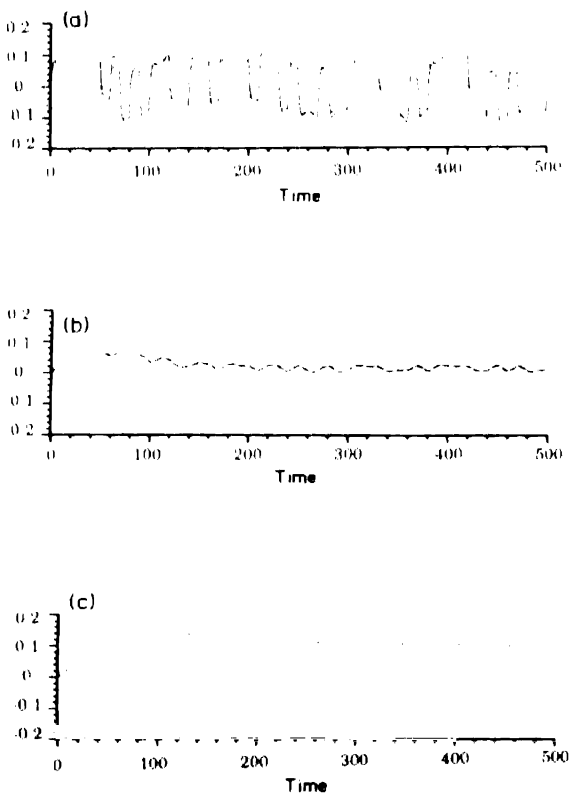


FIG. 10. Simulated responses of x_B and y_D to the unit step disturbances in z_F and F with 20% input uncertainty and pseudo-random binary measurement noise. (a) Measurement set tray #1/Tray #41; (b) Measurement set tray #7/Tray #35; (c) Measurement set tray #17/Tray #25; dashed line $\rightarrow x_B$; solid line $\rightarrow y_D$.

Acknowledgements—Support from the National Science Foundation and the Petroleum Research Fund administered by the American Chemical Society is gratefully acknowledged.

References

Bequette, B. W. and T. F. Edgar (1986). The equivalence of non-interacting control system design methods in distillation. *Proc. Amer. Control Conf.* Seattle, WA, 31–37.

Doyle, J. C. (1982). Analysis of feedback systems with structured uncertainties. *Proc. IEEE Pt. D.*, **129**, 242–247.

Doyle, J. C., J. E. Wall and G. Stein (1982). Performance and robustness analysis for structured uncertainty. *Proc. IEEE Conf. on Decision and Control*, Orlando, FL.

Doyle, J. C. (1984). Lecture notes in advances in multivariable control. *ONR/Honeywell Workshop*, Minneapolis, MN.

Francis, B. A. (1987). *A Course in H_2 Control Theory*. Springer, Heidelberg.

Harris, T. J., J. F. MacGregor and J. D. Wright (1980). Optimal sensor location with an application to a packed-bed tubular reactor. *AIChE J.*, **26**, 910–916.

Joseph, B. and C. B. Brosilow (1978). Inferential control of processes. *AIChE J.*, **24**, 485–509.

Kwakernaak, H. and R. Sivan (1972). *Linear Optimal Control Systems*. Wiley-Interscience, New York.

Kumar, S. and J. H. Seinfeld (1978a). Optimal location of measurements in tubular reactors. *Chem. Engng. Sci.*, **33**, 1507–1516.

Kumar, S. and J. H. Seinfeld (1978b). Optimal location of measurements for distributed parameter system. *IEEE Trans. Aut. Control*, **AC-23**, 690–698.

Lee, J. H. (1991). Robust inferential control: A methodology for control structure selection and inferential control system design in the presence of model/plant mismatch. Ph.D. Thesis, California Institute of Technology, Pasadena, CA.

Moore, C., J. Hackney and D. Canter (1987). Selecting sensor location and type for multivariable processes. *Shell Process Control Workshop*, Butterworth, Boston.

Morari, M. and G. Stephanopoulos (1980). Minimizing unobservability in inferential control schemes. *Int. J. Control*, **31**, 367–377.

Morari, M. and E. Zafiriou (1989). *Robust Process Control*. Prentice Hall, Englewood Cliffs, NJ.

Skogestad, S. and M. Morari (1988). Some new properties of the structured singular value. *IEEE Trans. Aut. Control*, **AC-33**, 1151–1154.

Skogestad, S., M. Morari and J. Doyle (1988b). Robust control of ill-conditioned plants. High-purity distillation. *IEEE Trans. Aut. Control*, **AC-33**, 1092–1105.

Weber, R. and C. Brosilow (1972). Use of secondary measurements to improve control. *AIChE J.*, **18**, 614–623.

Appendix

Definition of the Structured Singular Value The Structured Singular Value (μ) is defined as follows:

Definition A Let $M \in \mathbb{C}^{n \times m}$ and define the set Δ as follows

$$\Delta = \left\{ \tilde{\Delta} \in \mathbb{C}^{n \times m} : \tilde{\Delta} = \begin{bmatrix} \Delta_1 & & \\ & \ddots & \\ & & \Delta_k \end{bmatrix}, \right. \\ \left. \Delta_i \in \mathbb{C}^{m_i \times n_i}, \sum_{i=1}^k m_i = m, \sum_{i=1}^k n_i = n \right\} \quad (1)$$

Then $\mu_\Delta(M)$ (μ of M with respect to the uncertainty structure Δ) is defined as

$$\mu_\Delta(M) = \begin{cases} \left[\min_{\tilde{\Delta} \in \Delta} \left(\sigma(\tilde{\Delta}) \det(I + M\tilde{\Delta}) \right) \right]^{-1} \\ 0 \text{ if } \exists \text{ no } \tilde{\Delta} \in \Delta \text{ such that } \det(I + M\tilde{\Delta}) = 0. \end{cases} \quad (2)$$

Derivation of the Hankel norm condition We derive the condition (11) in Section 2.3. One can easily obtain the following equality by using the Q -parametrization of all stabilizing controllers:

$$\min_{K \in \mathbb{K}_s} \|F_{12}(K, \Delta_d)\|_{\Delta_d} = \min_{Q \in RH_\infty} \|W_p(G_{1d} - G_{1m}Q)G_{2d}W_d\|_\infty \quad (3)$$

where RH_∞ denotes the set of all rational transfer function matrices analytic in the closed Right-Half-Plane (RHP). Suppose W_p and W_d were chosen so that $W_p, W_d \in RH_\infty$. Assuming that $W_p G_{1m}(j\omega)$ and $G_{2d}W_d(j\omega)$ have constant ranks for $0 < \omega < \infty$, we can easily perform the following inner-outer and coinner-coouter factorizations:

$$W_p G_{1m} = (W_p G_{1m})_i (W_p G_{1m})_o \quad (4)$$

$$G_{2d}W_d = (G_{2d}W_d)_i (G_{2d}W_d)_o \quad (5)$$

For strictly proper systems, the assumption of constant ranks for $0 < \omega < \infty$ does not hold, however, they can be approximated as proper systems up to arbitrarily high frequency.

Theorem A Assume that $W_p G_{1m}$ and $G_{2d}W_d$ are square [i.e. $\dim(c) = \dim(m)$ and $\dim(s) = \dim(d)$]. Also assume that $(W_p G_{1m})_i (W_p G_{1d}W_d)(G_{2d}W_d)_i$ exists and is analytic in $j\omega$ -axis where $(\cdot)_i$ denotes the adjoint operator [i.e. $M(s) = M^T(-s)$]. Then

$$\min_{K \in \mathbb{K}_s} \|F_{12}(K, \Delta_d)\|_{\Delta_d} = 1 \quad (6)$$

if and only if

$$\|[(W_p G_{1m})_i (W_p G_{1d}W_d)(G_{2d}W_d)_i]\|_H = 1 \quad (7)$$

where $[\cdot]_i$ denotes the antistable factor and $\|\cdot\|_H$ denotes the Hankel norm.

Proof Straightforward from Nehari's Theorem (Francis, 1987).

Reduced Order Process Modelling in Self-tuning Control*

C. C. HANG†‡ and D. CHIN†

Key Words—Model based self-tuning control; reduced order model; deadtime; frequency response

Abstract—One major objection to the model-based auto-tuning and self-tuning control is based on the observation by Bristol that a reduced order model being identified would not yield good control of the actual high order process. In this paper we have re-examined the work of Bristol and argued that this observations could not be generalized. We then consider reduced order modelling which incorporates a deadtime element. It is shown that the deadtime could help to provide a close match of the frequency response near the critical frequencies. Extensive simulation studies have also confirmed that reduced order modelling with deadtime included could be an adequate model for practical self-tuning control.

1. Introduction

AN ESSENTIAL part of the self-tuning control involves parameter estimation of an explicit or implicit process model (Åström, *et al.*, 1977; Åström and Wittenmark, 1984). In order to reduce the number of parameters to be estimated for practical on-line self-tuning control, the order of the process model chosen is often smaller than that of the actual process hence introducing a potential mismodelling problem. This was highlighted by Bristol in conjunction with his work on a pattern recognition based adaptive control concept (Bristol, 1970, 1977, 1988). The study of mismodelling addresses principally the process structure errors, which are different from parametric errors. Bristol conjectured and demonstrated successfully through some examples that a reduced order model so identified would not yield good process control. Upon this premise, he objected to the concept of self-tuning control based on process model and instead developed the now well-known and commercially successful pattern recognition-based EXACT adaptive controller.

In this paper we have re-examined the problem of mismodelling. We shall put forward the argument that though Bristol's work and conclusions on mismodelling were relevant and correct, they could not be generalized, particularly when a deadtime element is included. It is inspired and supported by a study of frequency response matching as proposed by Wahlberg and Ljung (1986). The paper is organised as follows. Section 2 summarizes Bristol's work and questions the generality of the results and conclusions drawn. Section 3 presents the study of frequency response matching and simulation results when a deadtime element is included in reduced order modelling. Section 4 discusses the relevance of this study to adaptive control and gives the conclusions drawn.

2. The mismodelling experiments

The details of the mismodelling experiments conducted by Bristol were given elsewhere (Bristol, 1970, 1977). The experiments compared the closed-loop step responses between a high order process controlled by a PI controller and a reduced, second order model identically controlled. The model was derived in two ways, open-loop and closed-loop step response matching.

In essence, Bristol observed that the model derived under open-loop identification gave unacceptable control performance when the loop was closed around the process. Likewise the model derived under closed-loop identification gave grossly different open-loop responses when compared with the original process. However, a convergence to satisfactory control performance could be obtained in the case of closed-loop identification after several iterations. But for the case of the self-tuning regulators, such a convergence to acceptable control performance could not be assured since the self-tuning regulator structure uses an equivalent of open-loop identification structure (Bristol, 1977, 1988).

It is noted that Bristol's experiments utilized step responses as the basis for modelling—in both open-loop and closed-loop identification. Such an approach may not be appropriate as it may not sufficiently excite the relevant natural modes of the process for the purpose of system identification (Åström, *et al.*, 1977). Furthermore, the criterion used in the identification process was the integral square error of the step response; this means that the identification was essentially skewed towards the steady state portion of the set of data points. Therefore, the matching would favour the low frequency region of the process's frequency response, again increasing the possibility of mismodelling at the critical frequencies (Wahlberg and Ljung, 1986). Finally, since the matching process involved only step responses, the results emphasised a magnitude match. There was little consideration for the phase of the process and no attempt to match the model and process phase frequency responses.

From the above discussion, it is not certain that Bristol's observations could be generalized to completely reject the concept of self-tuning control with reduced-order modelling. It has been shown recently by Wahlberg and Ljung (1986) that frequency response matching near the critical point is more appropriate for assessing the accuracy of the reduced order modelling for the purpose of stable feedback control. In the case of least squares parameter estimation, this can be achieved by suitable choice of data filtering to focus on the frequency range of importance. In the following section, we shall attempt another approach involving the use of a deadtime element to address this mismodelling problem.

3. Reduced order modelling with deadtime

In order to facilitate easy comparison, the experiments were conducted on the same process as used by Bristol (Bristol, 1970, 1977):

$$G(s) = \frac{1}{(1 + 0.62s)(1 + 17.2s)} \quad (1)$$

* Received 16 June 1989; revised 12 July, 1990; received in final form 29 August 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor L. Valavani under the direction of Editor P. C. Parks.

† Department of Electrical Engineering, National University of Singapore, Kent Ridge, Singapore.

‡ Author to whom all correspondence should be addressed.

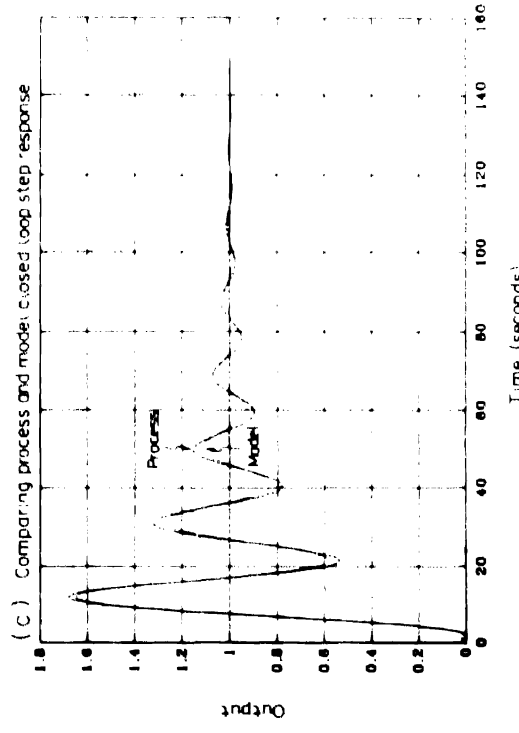
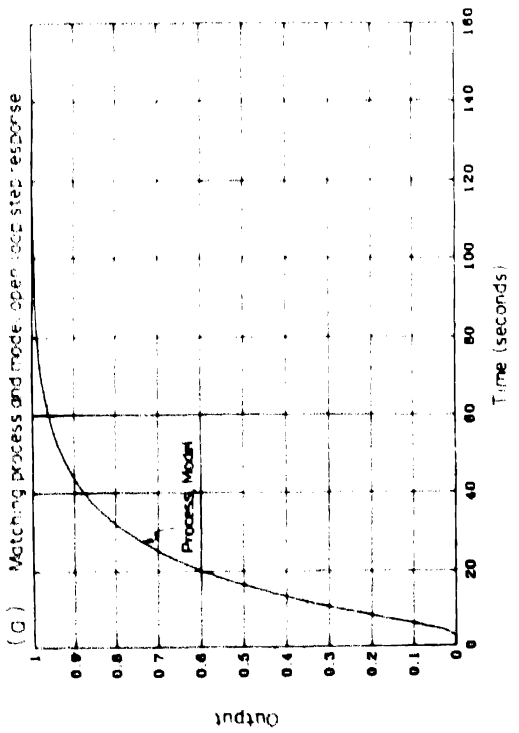
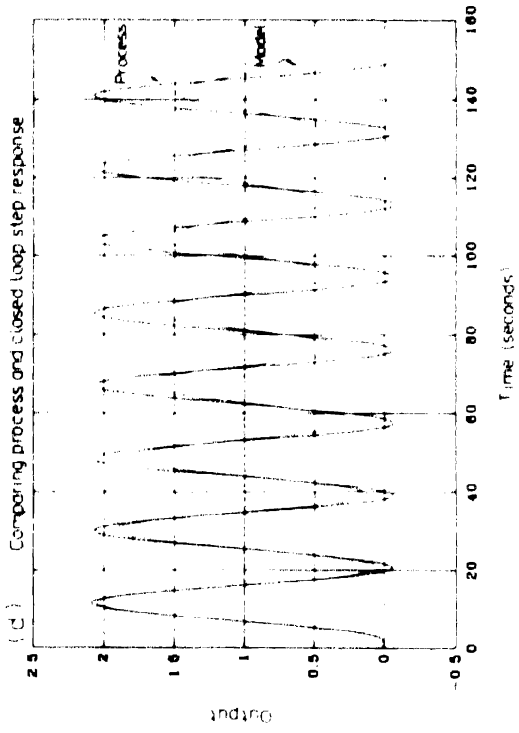
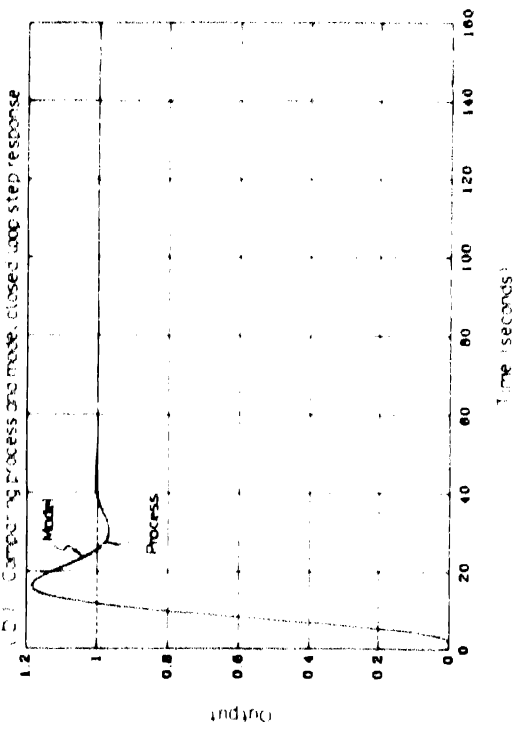


Fig. 1. Open loop and closed loop responses of process and model

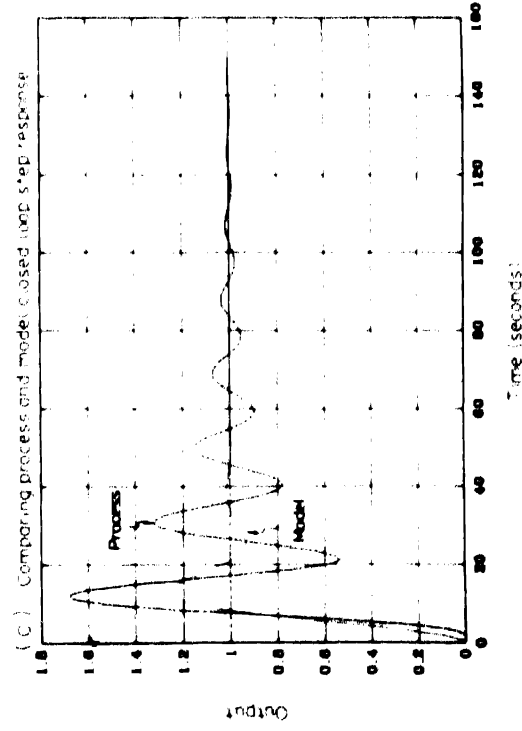
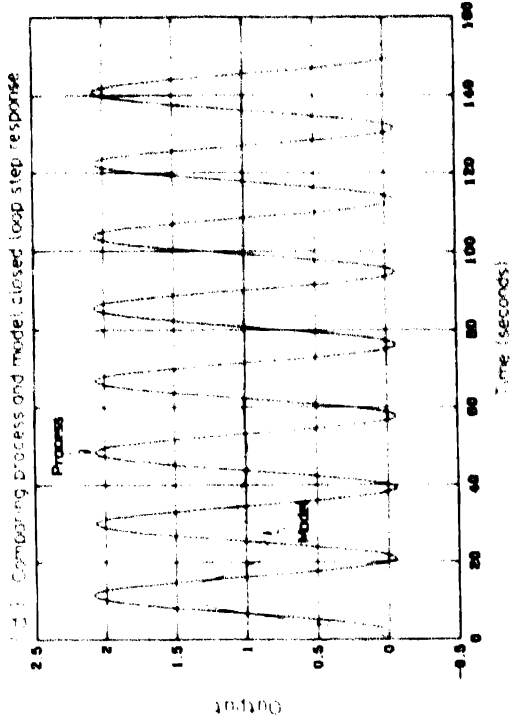
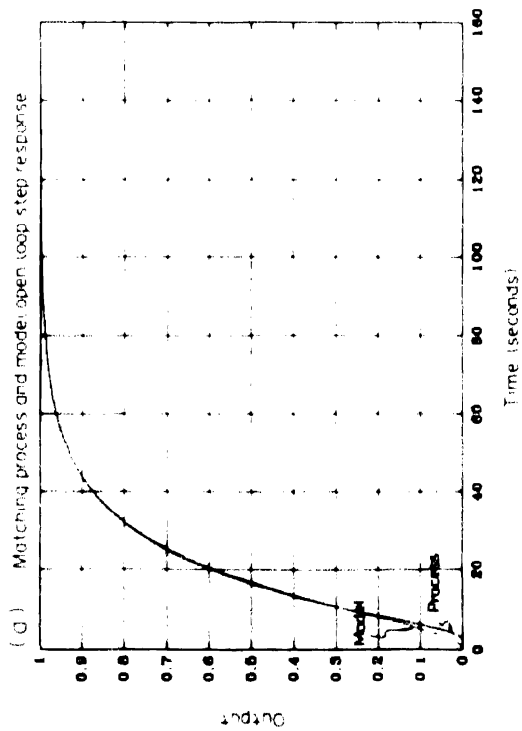
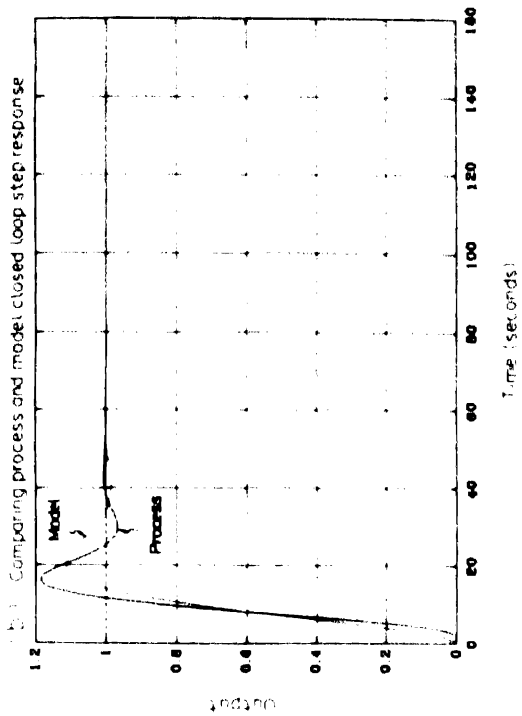


Fig. 2. Open-loop and closed-loop step responses of process and model B (without deadline)

Inspired by Wahlberg and Ljung (1986), we shall approach the study of reduced order modelling from a frequency domain perspective. First, we shall consider continuous models identified by matching the time responses, similar to the approach used by Bristol except that a pseudo-random binary sequence (PRBS) test signal is used instead of a step input. Two models were identified. The model, *A*, assumed to be second-order with deadline, was identified to be:

$$G_M(s) = \frac{e^{-2.4s}}{(1 + 1.98s)(1 + 17.11s)}$$

(2)

For comparison, a second model, *B*, assumed to be second-order but without deadline as in the Bristol's experiment, was identified to have time constants of 4.8 and 16.7. The criterion used for evaluating the goodness of fit for

both models was the integral square error between the process and model open-loop responses.

Consider Fig. 1(a) which gives the open-loop step responses of the process and model *A*, and Fig. 2(a) the same for the process and model *B*. Note the closeness of match of the two models to the process. From the process open-loop response alone, one might have accepted the plausibility of representing the process by a second order model without deadline, if the process was not known to have a higher order structure. However, when the loop was closed around the process and the models, each identically controlled by a proportional-integral (PI) controller, deviations became apparent. These deviations are obvious from Figs 1 and 2 which give the closed-loop responses for three sets of controller settings (proportional gain *K* and integral time *T_i*) as tabulated in Table 1. These settings were chosen to operate the closed-loop process near optimal, with poor damping and near unstable modes respectively in order to examine if these characteristics could be reflected in the models, as suggested by Bristol (1970, 1977)

The discrepancy between the open-loop and closed-loop responses for model *B* is evident from the plots, whereas model *A* can be seen to represent the process sufficiently. To understand the underlying disparity, we examine the Nyquist plots of the process and models as given in Fig. 3. This is

TABLE 1. PI CONTROLLER SETTINGS

Controller settings	<i>K</i>	<i>T_i</i>
Figs 1(b), 2(b)	2.8	17
Figs 1(c), 2(c)	5.6	17
Figs 1(d), 2(d)	6.63	10

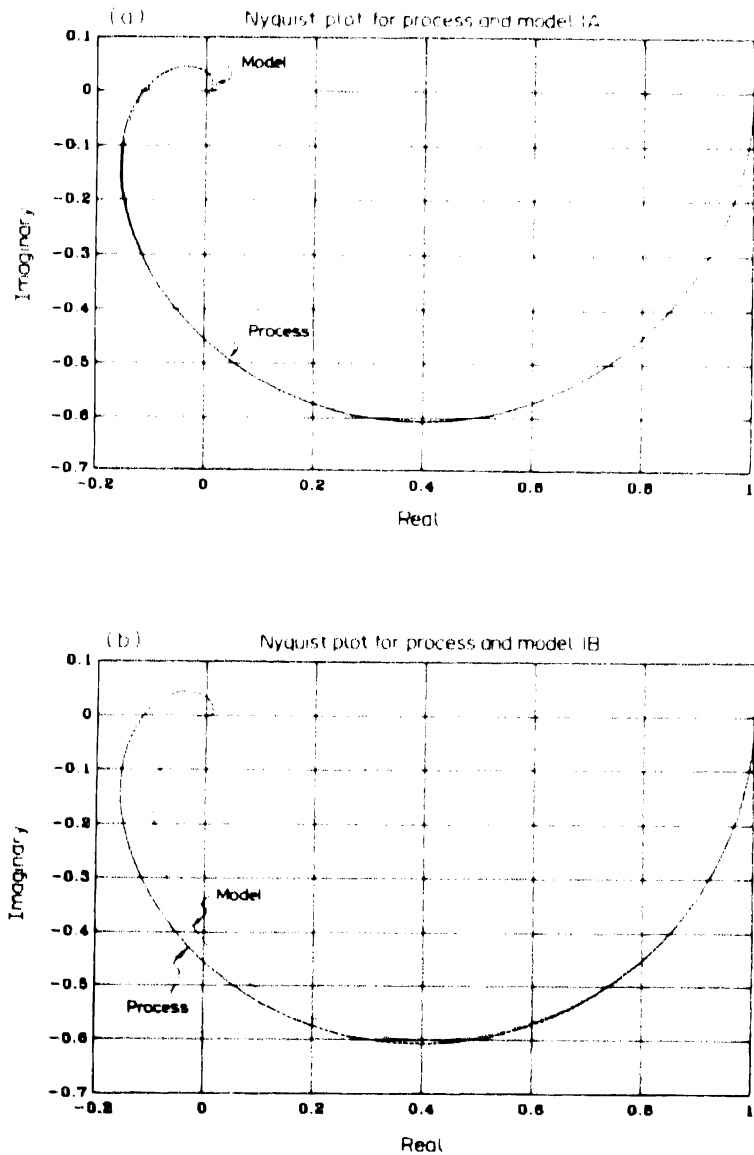


FIG. 3. Frequency response of process, models *A* and *B*.

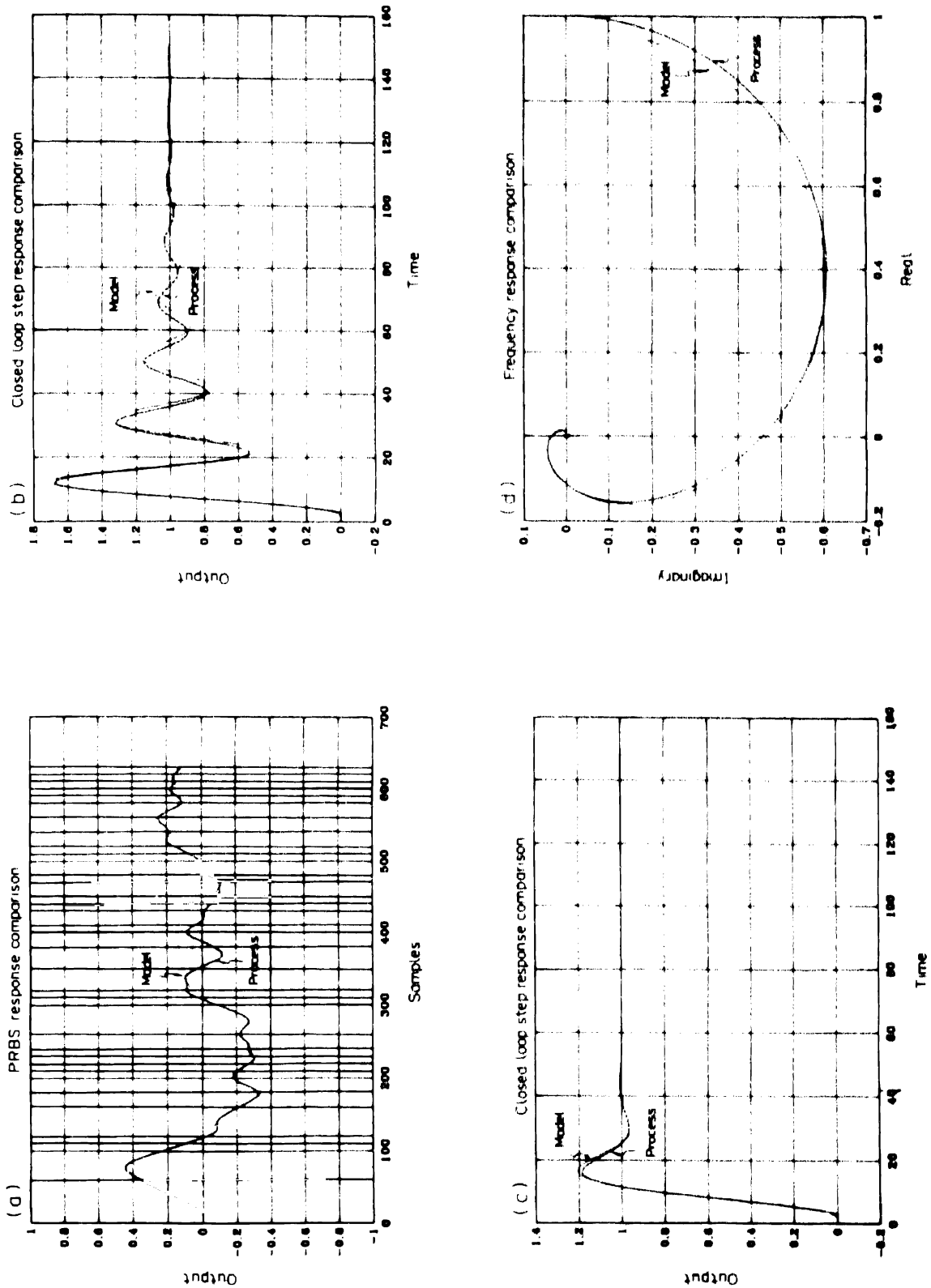


FIG. 4. Results of discrete time model C.

where the crux of the mismodelling argument resides. In model *B*, the open-loop response matching did not yield a good match of the frequency response at higher frequencies. But the inclusion of a suitable deadtime element in model *A* provided a chance not only to match the magnitude but also the phase frequency response of the model to the process resulting in a close match in the Nyquist plot between the process and model *A*.

This observation affirmed two points stated earlier. Firstly, open-loop step responses are poor guides for modelling. They may be used as supplementary guides in the matching process but cannot be the sole basis to arrive at a reduced order model. If anything at all, the frequency response of the process should form the basis for model identification or verification. Secondly, an open-loop step response favours a low frequency identification as evident from the Nyquist plot of model *B*. However, it is the higher frequency end of the process frequency response, especially the region around its Nyquist frequency, that is of significance in closed-loop control (Wahlberg and Ljung, 1986). This is because closing the loop around the process will create its resonant closed-loop modes that must be modelled. Bristol's observation on mismodelling was due largely to these unmodelled modes.

But for model *A*, the phase lag introduced by the deadtime overcame this problem of mismodelling. It could be seen from Figs 1(c) and (d) that even in the limiting cases, the model's responses corresponded very closely with those of the process. This points in the direction that a reduced order model with deadtime can be an adequate representation of the original process and hence a basis for control analysis and design.

As practical self-tuning control is invariably implemented in digital form, a discrete time model, *C*, was also identified using a PRBS excitation signal. The closed-loop responses of the model and process, given in Fig. 4 had the same controller settings as the continuous time case, by means of direct discretization. The simulations confirmed the applicability of the model in the discrete time domain. This point is evident both from the frequency response matching and the closed-loop responses of the model and the process.

Note from the closed-loop response that model *C* was nonminimum phase. This came about because the deadtime approximated was lower than what the process actually exhibited. From extensive simulation experience, it has been found that as long as the majority of the deadtime has been explicitly modelled, the residual deadtime can be safely reflected as nonminimum phase zero (De Souza *et al.*, 1988) which is automatically identified by the estimator. In discrete identification, variable deadtime can further be identified by overparameterizing the numerator polynomial of the process model.

4. Conclusions

In general, no practical process can be precisely modelled implying, then, all modelling effort is inherently faulty. Therefore, techniques considered should reduce some

mismodelling to a minimum. Even with exact knowledge of the process order, a full-order parameter estimator may not be practical due to consideration of the speed of parametric convergence, computational efficiency and precision. On the other hand, a reduced-order model requires only a fixed and small number of parametric estimates. This simpler model not only improves the convergence of the estimates but may also avoid the explosion of sensitive internal process identification data. In the example cited, the reduced-order model has only 3 parameters compared to 8 of the original process. It has been illustrated that the inclusion of a deadtime in the model may be sufficient to compensate for the gross misrepresentation of the process order. This augments the technique of data filtering (Wahlberg and Ljung, 1986) to improve the estimation accuracy near the critical frequencies for closed-loop stability. Thus, the problem of mismodelling which Bristol highlighted becomes much less significant.

The results have direct relevance and application in the area of auto-tuning and self-tuning control. This is because the modelling structure that was employed in the simulation studies is equivalent to that used for auto-tuning and self-tuning, as both essentially make use of open-loop identification.

The main conclusions can be summarized as follows:

- (a) Bristol's results based on step-response matching could not be generalized. Frequency response matching near the critical frequencies should instead be used to judge the adequacy of reduced order modelling.
- (b) Reduced order modelling with deadtime can compensate for the phase lag disparity between the process and model. This suitably augments the technique of data filtering to ensure that the estimated model could be an adequate basis for the purpose of controller design.

References

- Åström, K. J. and B. Wittenmark (1984). *Computer Controlled Systems: Theory and Design*. Prentice-Hall, Englewood Cliffs, NJ.
- Åström, K. J., U. Borisson, L. Ljung and B. Wittenmark (1977). Theory and application of self-tuning regulators. *Automatica*, **13**, 457-476.
- Bristol, E. H. (1970). Adaptive control odyssey. *Proc. ISA Annual Conf.* Philadelphia, pp. 561-70.
- Bristol, E. H. (1977). Pattern recognition: an alternative to parameter identification in adaptive control. *Automatica*, **13**, 197-202.
- Bristol, E. H. (1988). *The EXACT Pattern Recognition Adaptive Controller: A User-oriented Commercial Success*. Foxboro Company, Massachusetts.
- De Souza, C. E., G. C. Goodwin, D. Q. Mayne, and M. Palaniswami (1988). An adaptive control algorithm for linear systems having unknown time delay. *Automatica*, **24**, 327-341.
- Wahlberg, B. and L. Ljung (1986). Design variables for bias distribution in transfer function estimation. *IEEE Trans. Aut. Control*, **AC-31**, 134-144.

A Sensitivity Approach to Optimal Spline Robot Trajectories*

A. DE LUCA,†‡ L. LANARI† and G. ORIOLO†

Key Words—Robots; splines; optimization; nonlinear programming; sensitivity analysis; trajectory planning.

Abstract—A robot trajectory planning problem is considered. Using smooth interpolating cubic splines as joint space trajectories, the path is parameterized in terms of time intervals between knots. A minimum time optimization problem is formulated under maximum torque and velocity constraints, and is solved by means of a first order derivative-type algorithm for semi-infinite nonlinear programming. Feasible directions in the parameter space are generated using sensitivity coefficients of the active constraints. Numerical simulations are reported for a two-link Scara robot. The proposed approach can be used for optimizing more general objective functions under different types of constraints.

Introduction

OPTIMAL TRAJECTORY planning may considerably improve robot performance in industrial applications, particularly when productivity rate or energy consumption are of primary concern. In order to provide true optimal motion under actuator limitations, the full nonlinear manipulator dynamics has to be explicitly considered in the trajectory planning phase. The interactions between geometric, kinematic and dynamic issues substantially increase problem complexity with respect to purely kinematic approaches.

Specific classes of optimal robot motion planning problems have been recently solved, with minimum time as objective and torque limits as constraints. When the task is a point-to-point motion, a number of numerical approaches are available to minimize the traveling time, e.g. a modified gradient-type algorithm (Weinreb and Bryson, 1985), the multiple-shooting technique for nonlinear TPBVP (Geering *et al.*, 1986), and a dynamic programming scheme (Sahar and Hollerbach, 1986). All the above solution methods are computationally intensive. Moreover, the introduction of state constraints—like joint limits or maximum velocity bounds—brings in additional complexity. In any case, the main limitation of point-to-point motion planning is the unpredictability of the obtained path, which can be dangerous in presence of obstacles.

Alternatively, the robot may be required to follow a safe prespecified geometric path joining the initial to the final

point, either in joint or in cartesian space. Assuming that a continuous parameterization can be given for the whole path, Bobrow *et al.* (1985) and Shin and McKay (1985) have derived an efficient solution algorithm for the minimum time problem, directly working in the parameter phase-plane. It should be emphasized that the efficiency of their algorithm strongly relies on the particular form of the cost criterion. Also, the *a priori* specification of an overall geometric path and its continuous parameterization are requirements which may be too restrictive or cumbersome for real applications.

Most commonly, the task planner provides the trajectory planner with a robotic task description which is intermediate between the above two. Especially in complex environments, the typical output of the task planner is a sequence of cartesian poses (i.e. positions and orientations) for the end-effector, which have to be interpolated. In principle, intermediate poses are not restricted, but a safe overall path can be guaranteed by increasing the number of specified poses in proximity of obstacles. The problem faced here is how the trajectory planner can perform this interpolation in an optimal way. The class of interpolating functions should be chosen so to give nice smoothness properties, thus avoiding excitation of the mechanical structure, together with a low curvature profile.

The most appealing class of functions for generating robotic paths that satisfy the above specifications are spline functions, which are piecewise cubic polynomials smoothly interpolating a sequence of knots (de Boor, 1978). Splines with continuous second derivative (C^2 -splines) have been widely used in robotic applications, e.g. to obtain minimum time trajectories under purely kinematic constraints (Lin *et al.*, 1983). The minimization algorithm was the Nelder-Mead flexible polyhedron search, being based only on function evaluations, it has slow convergence and may stop in false constrained minima. Bobrow (1988) used C^1 -splines for approximating point-to-point minimum time paths in presence of obstacles.

In this paper, a new method is presented for planning smooth optimal robot trajectories interpolating a given sequence of points. The trajectory is a C^2 -spline passing through n knots, with boundary conditions on initial and final velocity. The $n - 1$ time intervals between the knots uniquely parameterize the path. A minimum time problem will be considered here, with both maximum torque and velocity limits. It turns out to be an optimization problem with infinite-dimensional constraints, which is solved via an efficient algorithm proposed by Gonzaga *et al.* (1980). This requires computation of the gradients of the active constraints, namely the sensitivity functions of the constraints with respect to variations of the design parameters. As an intermediate step, it is necessary to compute the sensitivity of the spline functions, which has its own interest and may be relevant also for other applications.

It must be stressed that the approach proposed here is conceptually different from the most common one, that would require *first* to build a spline interpolating the knots, and *then* find the optimal time history on this assigned path, using the algorithm of Bobrow *et al.* (1985). In fact, the

* Received 24 February 1989; revised 10 July 1990; received in final form 25 August 1990. The original version of this paper was presented at the IFAC Symposium on Robot Control (SYROCO '88) which was held in Karlsruhe, Germany during October 1988. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak.

† Dipartimento di Informatica e Sistemistica, Università degli Studi di Roma "La Sapienza", Via Eudossiana 18, 00184 Rome, Italy.

‡ Author to whom all correspondence should be addressed.

following aspects characterize the present method:

- The discrete parameterization allows one to handle any cost function of interest (e.g. energy consumption), while well-established exact solutions exist only in the minimum time case;
- Since the interpolating spline is a function of time intervals between knots, the resulting cartesian path changes throughout the optimization process, possibly yielding a shorter final traveling time than the one achievable on the "first-guess" path;
- Velocity bounds are explicitly considered; moreover, smoothness up to the second derivative is guaranteed, avoiding torque discontinuities.

After restating the essential steps for the generation of an interpolating spline trajectory, the minimum time problem is formulated as a semi-infinite nonlinear programming problem and solved by the algorithm of Gonzaga *et al.* (1980). To this goal, the sensitivity analysis of spline and torque functions with respect to time intervals will be derived. Numerical results obtained for two planar robot arms are presented. In the conclusions, some possible extensions of the proposed approach are outlined.

Spline trajectory generation

For a robot with N joints, let the task be assigned by a sequence of n cartesian poses P_1, P_2, \dots, P_n to be assumed by the end-effector at unspecified time instants t_1, t_2, \dots, t_n . Initial and final cartesian velocities are given. Using inverse kinematics, these data are transformed into joint configurations $q_{j1}, q_{j2}, \dots, q_{jn}$, and into initial and final joint velocities v_{j1} and v_{jn} , with $j = 1, \dots, N$. Let $h_i = t_{i+1} - t_i$, $i = 1, \dots, n-1$, be the time intervals between knots, and $\mathbf{h} = [h_1, \dots, h_{n-1}]^T$. For a generic joint j , a trajectory is obtained by interpolation using a C^2 -spline $Q_j(t)$ —a piecewise cubic polynomial. Denote by ω_{ji} the value of the j th joint acceleration at the i th knot. Each of the $n-1$ cubics $Q_{ji}(t)$ constituting the spline can be written in terms of the ω_{ji} (Lin *et al.*, 1983) as

$$Q_{ji}(t) = \frac{(t_{i+1} - t)^3}{6h_i} \omega_{ji} + \frac{(t - t_i)^3}{6h_i} \omega_{ji+1} + \left[\frac{q_{ji+1} - h_i \omega_{ji+1}}{h_i} \right] (t - t_i) + \left[\frac{q_{ji} - h_i \omega_{ji}}{h_i} \right] (t_{i+1} - t) \quad (1)$$

where $t \in [t_i, t_{i+1}]$. Spline velocity $\dot{Q}_j(t)$ and acceleration $\ddot{Q}_j(t)$ are piecewise quadratic and linear functions, respectively. The continuity requirement for the acceleration at the internal knots is automatically satisfied, since $\ddot{Q}_{j,i-1}(t_i) = \ddot{Q}_{j,i}(t_i) = \omega_{ji}$. Given a timing for the trajectory (i.e. a vector \mathbf{h}), the interpolating spline is completely specified by (1) once knots accelerations ω_{ji} are computed. Denoting by Ω_j the vector of knots accelerations for the j th joint, and imposing the continuity for velocities at the internal knots as well as boundary conditions, for each joint a tridiagonal linear system is derived in the form

$$\mathbf{A}\Omega_j = \mathbf{b}_j, \quad j = 1, \dots, N, \quad (2)$$

where

$$\mathbf{A} = \begin{bmatrix} 2h_1 & h_1 & & & \\ h_1 & 2(h_1 + h_2) & & & \\ & & \ddots & & \\ & & & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ & & & h_{n-1} & 2h_{n-1} \end{bmatrix}$$

$$\mathbf{b}_j = \begin{bmatrix} 6\left(\frac{q_{j2} - q_{j1}}{h_1} - v_{j1}\right) \\ 6\left(\frac{q_{j3} - q_{j2}}{h_2} - \frac{q_{j2} - q_{j1}}{h_1}\right) \\ \vdots \\ 6\left(\frac{q_{jn} - q_{j,n-1}}{h_{n-1}} - \frac{q_{j,n-1} - q_{j,n-2}}{h_{n-2}}\right) \\ 6\left(v_{jn} - \frac{q_{jn} - q_{j,n-1}}{h_{n-1}}\right) \end{bmatrix}$$

For $\mathbf{h} > \mathbf{0}$, matrix \mathbf{A} is diagonally dominant and the solution of (2) is unique and continuous in the elements of \mathbf{h} . Note that \mathbf{A} is the same for all joints. Solution Ω_j of (2) is efficiently obtained without the need of pivoting by the following Gauss-type algorithm (Stoer and Burlirsch, 1980):

$$\omega_{jn} = \dot{\omega}_{jn}, \quad \omega_{ji} = \dot{\omega}_{ji} - K_i \omega_{j,i+1}, \quad i = n-1, \dots, 1, \quad (3)$$

with

$$\dot{\omega}_{j1} = \frac{b_{j1}}{a_{j1}}, \quad \dot{\omega}_{ji} = \frac{b_{ji} - \dot{\omega}_{j,i+1} a_{j,i+1}}{a_{ji} - K_{i+1} a_{j,i+1}}, \quad i = 2, \dots, n, \quad (4)$$

$$K_i = \frac{a_{j,i+1}}{a_{ji}}, \quad K_i = \frac{a_{j,i+1}}{a_{ji} - K_{i+1} a_{j,i+1}}, \quad i = 2, \dots, n-1$$

where a_{hk} are the elements of \mathbf{A} , and b_{hk} of \mathbf{b}_j .

For any choice of parameter vector \mathbf{h} , a unique spline $Q_j(t, \mathbf{h})$ is built for each joint, forming a vector function $\mathbf{Q}(t, \mathbf{h})$. It should be noted that changing one or more of the parameters h_i will modify the whole interpolating spline. This is true unless a uniform scaling is performed on all the components of \mathbf{h} (Hollerbach, 1984).

The optimization problem

Each spline trajectory $Q_j(t, \mathbf{h})$ is parameterized in terms of vector \mathbf{h} . The optimal value of this vector can be determined according to a specified criterion and subject to proper constraints. If a minimum time problem is considered, typical constraints are symmetric velocity and torque limits for each joint, V_j and U_j respectively. Some peculiar features of the problem are exploited to obtain a compact formulation.

Velocity constraints. Since spline velocity is a piecewise quadratic function, its maximum value in any subinterval is attained either at the knot instants t_i, t_{i+1} , or at one intermediate instant $t_i^* \in [t_i, t_{i+1}]$ where $\dot{Q}_j(t_i^*, \mathbf{h}) = 0$. This zero-acceleration instant exists iff $\omega_{ji} \omega_{j,i+1} < 0$ holds, in which case $t_i^* = t_i + h_i \omega_{ji} / (\omega_{ji} - \omega_{j,i+1})$ and

$$\dot{Q}_j(t_i^*) = -\frac{h_i}{2(\omega_{j,i+1} - \omega_{ji})} \omega_{ji} \omega_{j,i+1} + \frac{q_{j,i+1} - q_{ji} - h_i(\omega_{j,i+1} - \omega_{ji})}{h_i} \quad (5)$$

Torque constraints. The maximum torque value can be attained anywhere along the spline trajectory; this constraint should be checked in all path points using robot dynamics

$$\mathbf{u}(t) = \mathbf{M}(\mathbf{q}(t))\ddot{\mathbf{q}}(t) + \mathbf{c}(\mathbf{q}(t), \dot{\mathbf{q}}(t)) + \mathbf{e}(\mathbf{q}(t)), \quad (6)$$

where $\mathbf{q} \in \mathbb{R}^N$ are the joint coordinates, \mathbf{u} are the actuator torques, \mathbf{M} is the inertia matrix, \mathbf{c} are the Coriolis and centrifugal forces, and \mathbf{e} is the gravity term. The torque $u_j(t, \mathbf{h})$ arising at joint j during motion along a spline trajectory is obtained by setting in (6) $\mathbf{q}(t) = \mathbf{Q}(t, \mathbf{h})$, together with its time derivatives.

Time interval constraints. Since parameterized splines are defined only for $\mathbf{h} > \mathbf{0}$, this constraint should be included in the optimization problem. The strict positivity requirement can be replaced by a lower bound derived from velocity limits. In fact, since $|q_{j,i+1} - q_{ji}|/h_i \leq V_j$ must hold on any subinterval, \mathbf{h} should satisfy the constraint

$$h_i \geq \kappa_i > 0 \quad \text{where} \quad \kappa_i = \max_{j=1, \dots, N} \left\{ \frac{|q_{j,i+1} - q_{ji}|}{V_j} \right\},$$

$$i = 1, \dots, n-1.$$

As a result, the following optimization problem is formulated:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n-1} h_i \\ \text{s.t.} \quad & |\dot{Q}_j(t_i, \mathbf{h})| \leq V_j \quad (i = 1, \dots, n), \\ & |\dot{Q}_j(t_i^*, \mathbf{h})| \leq V_j \quad (i = 1, \dots, n-1), \\ & \max_{t \in [t_i, t_{i+1}]} \{ |u_j(t, \mathbf{h})| \} \leq U_j, \quad j = 1, \dots, N, \\ & \mathbf{w} - \mathbf{h} \leq \mathbf{0}. \end{aligned}$$

This problem is an instance of the semi-infinite nonlinear

programming class

$$\min f(\mathbf{h})$$

$$\text{s.t. } g^l(\mathbf{h}) \leq 0, \quad l = 1, \dots, p,$$

$$\max_{\tau \in T} \phi'(\tau, \mathbf{h}) \leq 0, \quad j = 1, \dots, r,$$

where $\mathbf{h} \in \mathbb{R}^{n-1}$ is the vector of design parameters, g^l are conventional constraints (velocity constraints plus lower bounds on \mathbf{h}), while functional (infinite-dimensional) constraints are represented through the ϕ^j 's (torque constraints). The domain T over which the functional constraints have to be satisfied is $[t_i, t_n]$. In the minimum time problem, $f(\mathbf{h})$ and $g^l(\mathbf{h})$ are continuously differentiable, $\phi^j(\tau, \mathbf{h})$ are continuous in both arguments, while gradients $\nabla_{\mathbf{h}} \phi^j(\tau, \mathbf{h})$ are continuous w.r.t. \mathbf{h} .

For this class of problems, Gonzaga *et al.* (1980) have developed an efficient solution algorithm which is a combined phase I-phase II method of feasible directions. The method does not require an admissible starting point, and recovers feasibility in a finite number of iterations, already considering the objective function at this stage. At each iteration, a low-dimensional quadratic programming (QP) subproblem is solved to generate a search direction \mathbf{d} in the space of parameters \mathbf{h} . Directional derivatives of the objective function and of the ϵ -active (i.e. active or almost active) constraints are needed in this QP. The algorithm uses a proper discretization of the functional constraints, replacing the continuous domain T with a set T_s of mesh instants. This discretization must be tailored to the problem at hand. Since T is itself a function of the current \mathbf{h} , an adaptive strategy is devised to discretize T into T_s . At iteration q , the following set of mesh points is used.

$$T_s^{[q]} = \{t_{im}^{[q]} \mid t_{im}^{[q]} = t_i^{[q]} + \alpha_m h_i, \\ m = 0, 1, \dots, s, i = 1, \dots, n-1\} \quad (7)$$

where $\alpha_m = m/s \in [0, 1]$. In this way, each subinterval $[t_i, t_{i+1}]$ is uniformly sampled and the knots are always included in the discretization.

Sensitivity analysis

For the solution of the minimum time problem, the following directional derivatives are needed by the Gonzaga *et al.* (1980) algorithm

$$(\nabla f(\mathbf{h}), \mathbf{d}) = \sum_{i=1}^{n-1} d_i,$$

$$(\nabla g^l(\mathbf{h}), \mathbf{d}) = \pm \sum_{k=1}^n \frac{\partial Q_l(\tau, \mathbf{h})}{\partial h_k} d_k, \quad \text{with } \tau = t_i \text{ or } t_n^*,$$

$$(\nabla_{\mathbf{h}} \phi^j(\tau, \mathbf{h}), \mathbf{d}) = \pm \sum_{k=1}^n \frac{\partial u_j(\tau, \mathbf{h})}{\partial h_k} d_k, \quad \text{with } \tau = t_{im} \quad (8)$$

In (8), the evaluation of spline (namely $Q_i, \dot{Q}_i, \ddot{Q}_i$) sensitivity with respect to changes of the generic time interval h_k is required. As a first step, the sensitivity of the solution to (2) (i.e. of knots accelerations ω_n) w.r.t. variations of h_k has to be derived. Since

$$\mathbf{A}(\mathbf{h}) \frac{\partial \Omega_i}{\partial h_k} = \frac{\partial \mathbf{b}_i(\mathbf{h})}{\partial h_k} - \frac{\partial \mathbf{A}(\mathbf{h})}{\partial h_k} \Omega_i \triangleq \tilde{\mathbf{b}}_i^{(k)},$$

the sensitivity $\partial \Omega_i / \partial h_k$ is the solution of a linear system having the same tridiagonal coefficient matrix as in (2) with constant $\tilde{\mathbf{b}}_i^{(k)}$ in place of \mathbf{b}_i . Thus, it can be found using again the recursive algorithm (3). Letting $\omega_n^{(k)} \triangleq \partial \omega_n / \partial h_k$, these are obtained as

$$\omega_n^{(k)} = \tilde{\omega}_n^{(k)}, \quad \omega_n^{(k)} = \tilde{\omega}_n^{(k)} - K_{i, n-1} \tilde{\omega}_{n-1}^{(k)}, \quad i = n-1, \dots, 1,$$

with

$$\tilde{\omega}_1^{(k)} = \frac{\tilde{b}_1^{(k)}}{a_{11}}, \quad \tilde{\omega}_n^{(k)} = \frac{\tilde{b}_n^{(k)} - \tilde{\omega}_{n-1}^{(k)} a_{n, n-1}}{a_{nn} - K_{i, n-1} a_{i, n-1}}, \quad i = 2, \dots, n.$$

The elements of vector $\tilde{\mathbf{b}}_i^{(k)}$ are

$$\tilde{b}_{i, n}^{(k)} = 0, \quad i = 1, \dots, k-1, k+2, \dots, n,$$

$$\tilde{b}_{i, k}^{(k)} = -2\omega_n - \omega_{n, k+1} - \frac{6(q_{i, k+1} - q_{i, k})}{h_k^3},$$

$$\tilde{b}_{i, k+1}^{(k)} = -\omega_n - 2\omega_{n, k+1} + \frac{6(q_{i, k+1} - q_{i, k})}{h_k^3}.$$

A more explicit treatment of the above expressions is pursued in (De Luca *et al.*, 1988). The above analysis is used to compute the sensitivity of the j th ($j = 1, \dots, N$) spline (1) at a generic mesh point t_{im} ($i = 1, \dots, n-1; m = 0, \dots, s$) which is

$$\frac{\partial Q_n(t_{im})}{\partial h_k} = \frac{h_i^2}{6} [(1 - \alpha_m)^3 \omega_n^{(k)} + \alpha_m^3 \omega_{n, i+1}^{(k)} \\ - \alpha_m \omega_{n, i+1}^{(k)} - (1 - \alpha_m) \omega_n^{(k)}] \\ + \delta_{ik} \frac{h_i}{3} [(1 - \alpha_m)^3 \omega_n + \alpha_m^3 \omega_{n, i+1} \\ - \alpha_m \omega_{n, i+1} - (1 - \alpha_m) \omega_n]$$

for $k = 1, \dots, n-1$, δ_{ik} being the Kronecker delta. Similarly, the velocity sensitivity at t_{im} is

$$\frac{\partial \dot{Q}_n(t_{im})}{\partial h_k} = \frac{h_i}{2} [\alpha_m^2 \omega_{n, i+1}^{(k)} - (1 - \alpha_m)^2 \omega_n^{(k)} \\ - \frac{h_i}{6} (\omega_{n, i+1}^{(k)} - \omega_n^{(k)}) \\ + \delta_{ik} \left[\frac{\alpha_m^2 \omega_{n, i+1} - (1 - \alpha_m)^2 \omega_n}{2} \right. \\ \left. - \frac{q_{i, i+1} - q_n}{h_i^2} - \frac{\omega_{n, i+1} - \omega_n}{6} \right],$$

while at the zero-acceleration instants one has

$$\frac{\partial \dot{Q}_n(t_n^*)}{\partial h_k} = \frac{h_i}{6} (\omega_{n, i+1}^{(k)} - \omega_n^{(k)}) \\ - \frac{h_i (\omega_{n, i+1}^2 \omega_n^{(k)} - \omega_n^2 \omega_{n, i+1}^{(k)})}{2(\omega_{n, i+1} - \omega_n)^2} \\ + \delta_{ik} \left[\frac{\omega_{n, i+1} - \omega_n}{6} + \frac{q_{i, i+1} - q_n}{h_i^2} \right. \\ \left. + \frac{\omega_n \omega_{n, i+1}}{2(\omega_{n, i+1} - \omega_n)} \right]$$

Finally, the sensitivity of spline acceleration is simply

$$\frac{\partial \ddot{Q}_n(t_{im})}{\partial h_k} = \alpha_m \omega_{n, i+1}^{(k)} + (1 - \alpha_m) \omega_n^{(k)}.$$

These quantities enter directly into the sensitivity of the functional constraints

$$\frac{\partial u_j}{\partial h_k} = \sum_{i=1}^N \left(\sum_{r=1}^N \frac{\partial m_{ji}}{\partial Q_i} \frac{\partial Q_i}{\partial h_k} \dot{Q}_i + m_{ji} \frac{\partial \dot{Q}_i}{\partial h_k} \right. \\ \left. + \frac{\partial c_i}{\partial Q_i} \frac{\partial Q_i}{\partial h_k} + \frac{\partial c_i}{\partial \dot{Q}_i} \frac{\partial \dot{Q}_i}{\partial h_k} + \frac{\partial c_i}{\partial \ddot{Q}_i} \frac{\partial \ddot{Q}_i}{\partial h_k} \right), \quad (9)$$

where m_{ji} , c_i and e_i are the elements of \mathbf{M} , \mathbf{c} and \mathbf{e} in (6). Note that in (9) derivatives of the dynamic model terms w.r.t. joint positions and velocities appear. These depend on the specific robot arm and can be computed automatically using symbolic manipulation languages (Neuman and Murray, 1984).

Numerical results

The proposed approach has been used to generate minimum time smooth spline trajectories for two different two-link SCARA-type robots. Programs were written in Fortran 77 on an AT personal computer and, at each iteration, the routine *E(MNAF)* of the NAG Workstation Library was used to solve the quadratic programming subproblem which provides the search direction. The dynamic model (6) of both arms takes on the explicit form

$$u_1 = (H_1 + 2H_2 \cos q_2) \ddot{q}_1 + (H_1 + H_2 \cos q_2) \ddot{q}_2 \\ - H_2(2\dot{q}_1 \dot{q}_2 + \dot{q}_2^2) \sin q_2 \\ u_2 = (H_1 + H_2 \cos q_2) \ddot{q}_1 + H_1 \ddot{q}_2 + H_2 \dot{q}_1^2 \sin q_2,$$

TABLE 1 PARAMETERS OF THE TWO ROBOTS USED IN THE EXAMPLES

	l_1 (m)	l_2 (m)	d_2 (m)	m_2 (kg)	m_p (kg)	J_1 (kg m ²)	J_2 (kg m ²)	J_p (kg m ²)
Example 1	0.5	0.5	0.25	1	0	0.084	0.084	0
Example 2	0.4	0.25	0.125	15	6	1.6	0.34	0.01

with

$$H_1 = J_1 + J_2 + J_p + m_2 l_1^2 + m_p (l_1^2 + l_2^2),$$
$$H_2 = m_2 l_1 d_2 + m_p l_1 l_2, \quad H_3 = J_2 + J_p + m_p l_2^2.$$

Note that $e(q) = 0$ since the motion is constrained on a horizontal plane. In the above equations m_i , l_i and J_i ($i = 1, 2$) are the mass, length and moment of inertia w.r.t. the axis of the driving joint for link i , while m_p and J_p are the mass and centroidal inertia of the payload. Also, d_2 is the distance between the axis of the second joint and the center of mass of the second link.

As a first example, a very light robot arm has been considered, whose parameters are reported in Table 1. The limit values are $V_1 = V_2 = 2 \text{ rad sec}^{-1}$ for the joint velocity, $U_1 = 7$, $U_2 = 2 \text{ Nm}$ for the torques. The robot task requires the arm to move along a C^2 -spline trajectory passing through a sequence of six joint knots (in rads): $\{q_1\} = \{0, 0.5, 0.75, 1, 1.25, 1.5\}$, $\{q_2\} = \{0, -0.5, -1, -1.5, -1, 0.5\}$, with zero initial and final velocity. The chosen initial design parameter is $h = [1, 0.5, 0.5, 0.5, 0.5]^T$, so that $t_n = 3 \text{ sec}$. On the resulting spline, velocity and torque of the second joint are both unfeasible, reaching the values of 4 rad sec^{-1} and -2.7 Nm . Figures 1-3 refer to the optimal solution found after 40 iterations. The optimal design parameter vector is $h^* = [0.37, 0.25, 0.34, 0.43, 1.07]^T$, from which $t_n^* = 2.46 \text{ sec}$. Both torques are saturated at the initial instant, while the second joint velocity saturates twice, near the second and fourth trajectory knots.

The obtained results deserve some comments. It is known that the minimum time torque profile along a parameterized trajectory is such that at each instant at least one actuator provides its maximum torque (Bobrow *et al.*, 1985). The

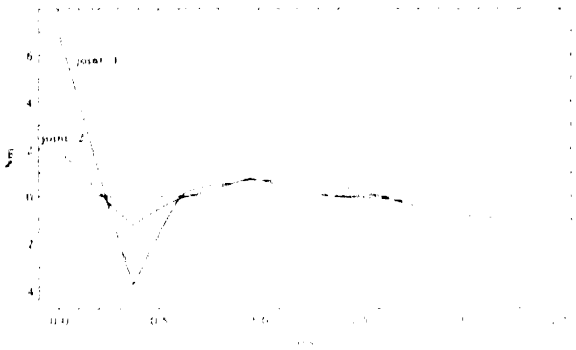


FIG. 3. Example 1. Optimal solution: torque.

reason for the absence of these saturated "flat tops" in the obtained profiles is twofold. First, the presence of velocity bounds prevents the torques from reaching their maximum values. Second, the requirement of a continuous acceleration excludes a bang-bang form for the torques.

As a second example, a planar motion of an IBM 7535 robot has been considered. Its parameters are shown in Table 1, while the bounds are $V_1 = V_2 = 1 \text{ rad sec}^{-1}$, and $U_1 = 9$, $U_2 = 25 \text{ Nm}$. The task is specified by a different sequence of six knots in the joint space: $\{q_1\} = \{q_2\} = \{0, 1, 0.2, 0.25, 0.3, 0.35, 0.4\}$ (rads), again with zero initial and final velocity. The initial design parameter is $h = [0.3, 0.3, 0.3, 0.3, 0.3]^T$, adding up to $t_n = 1.5 \text{ sec}$. The joint torques on the corresponding path are unfeasible, reaching the values of 45.6 Nm for the first joint and -12.2 Nm for the second. Feasibility is recovered after 5 iterations, with $t_n = 1.31 \text{ sec}$. Note that feasibility is recovered with a lower value of the objective function, as a result of the combined phase I-phase II algorithm. Figures 4-6 refer to the solution found after 34 iterations. The optimal design parameter is $h^* = [0.29, 0.07, 0.07, 0.08, 0.2]^T$, giving a traveling time $t_n^* = 0.71 \text{ sec}$.

It is interesting to note that the obtained torques approximate a bang-bang behavior, although such a profile is outside the C^2 -class. Two intermediate time intervals (i.e. h_2 and h_3) are in fact brought close to zero, while saturated torques are obtained in the first and in the last two intervals. These torque solutions are allowed since there is no velocity saturation in this case. If acceleration continuity is relaxed on our final path, the algorithm of Shin and McKay (1985) can be applied to obtain a lower minimum time solution.

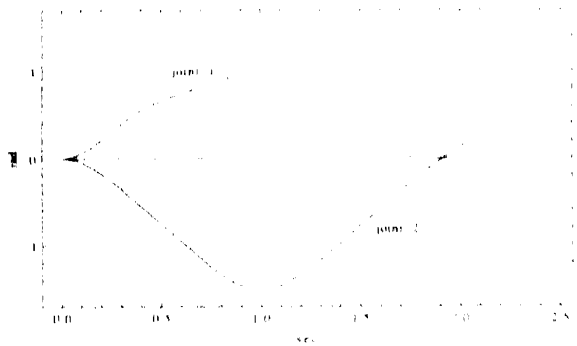


FIG. 1. Example 1. Optimal solution: position.

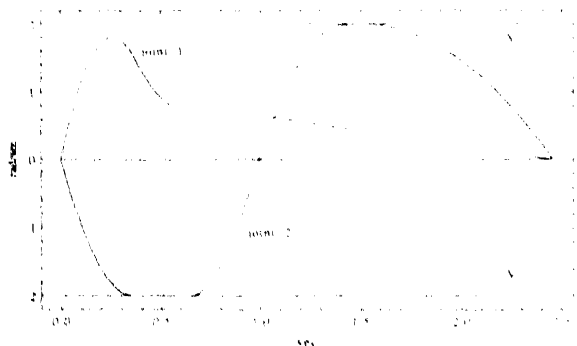


FIG. 2. Example 1. Optimal solution: velocity.



FIG. 4. Example 2. Optimal solution: position.

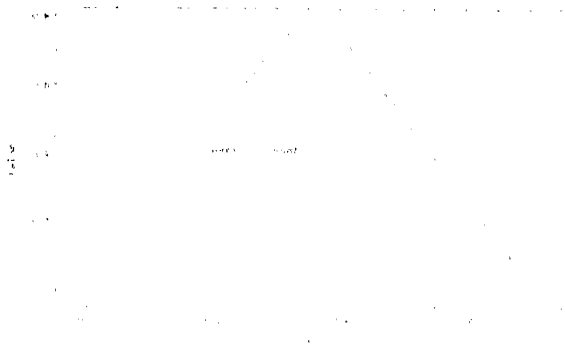


FIG. 5. Example 2. Optimal solution, velocity.

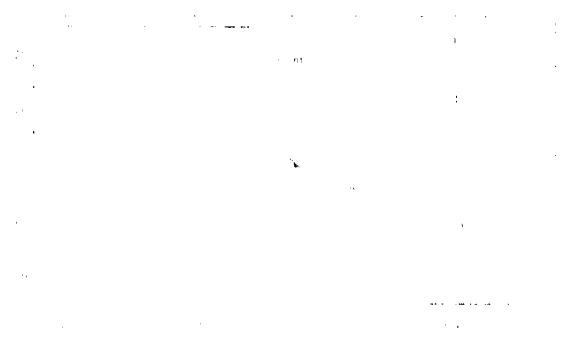


FIG. 6. Example 2. Optimal solution, torque.

However, only a negligible improvement is gained with respect to the proposed method, since $t_n^* = 0.707$ sec results. Other simulations, together with further details on the optimization algorithm, are reported in De Luca *et al.* (1988).

Conclusions

A new method has been presented for minimizing a general objective function along a spline trajectory under kinematic and dynamic constraints. In particular, minimization of the total traveling time along a path specified by a sequence of knots has been considered. The peculiarity of the proposed approach stands in the following aspects: (i) the problem is parameterized by a finite number of variables (i.e., the time intervals between the knots), thus allowing the use of an efficient nonlinear programming algorithm; (ii) general constraints on the robot state can be included directly in the formulation; (iii) the optimization is performed over the class of C^3 -splines, thus ensuring smooth generated trajectories, which are graceful for the mechanical structure of the robot. The inclusion of constraints on both velocity and torque, together with the continuity requirement assumed for the trajectory acceleration, produces interesting minimum time torque profiles.

The numerical optimization algorithm used is very robust and efficient, being based on gradient information. The sensitivity of the spline with respect to changes of the time

intervals between the knots has been explicitly derived. The obtained expressions can prove useful also in purely kinematic approaches to robot trajectory planning, as well as in other applications.

Use of the proposed method in optimization problems with more general objective functions under different types of constraints, like velocity-dependent torque bounds or joint limits, requires only little additional complexity. Nondifferentiable functions may also be treated, following the theoretical developments of the same basic algorithm as presented in (Polak *et al.*, 1983). This is of interest for tackling robot trajectory planning problems in the presence of obstacles, where nondifferentiable distance functions come into play (Gilbert and Johnson, 1985).

References

- Bobrow, J. E. (1988). Optimal robot path planning using the minimum-time criterion. *IEEE J. Robotics Aut.*, **RA-4**, 443-450.
- Bobrow, J. E., S. Dubowsky and J. S. Gibson (1985). Time-optimal control of robotic manipulators along specified paths. *Int. J. Robotics Res.*, **4**, 3-17.
- de Boor, C. (1978). *A Practical Guide to Splines*. Springer, Berlin.
- De Luca, A., L. Lanari and G. Oriolo (1988). Generation and computation of optimal smooth trajectories for robot arms. DIS Report 1388, Università di Roma "La Sapienza", Rome.
- Geering, H. P., L. Guzzella, S. A. R. Hepner and C. H. Onder (1986). Time-optimal motions of robots in assembly tasks. *IEEE Trans. Aut. Control*, **AC-31**, 512-518.
- Gilbert, E. G. and D. W. Johnson (1985). Distance functions and their application to robot path planning in the presence of obstacles. *IEEE J. Robotics Aut.*, **RA-1**, 21-30.
- Gonzaga, C., E. Polak and R. Trahan (1980). An improved algorithm for optimization problems with functional inequality constraints. *IEEE Trans. Aut. Control*, **AC-25**, 443-450.
- Hollerbach, J. M. (1984). Dynamic scaling of manipulator trajectories. *ASME J. Dyn. Syst. Meas. Control*, **106**, 102-106.
- Lin, C. S., P. R. Chang and J. Y. S. Luh (1983). Formulation and optimization of cubic polynomial joint trajectories for industrial robots. *IEEE Trans. Aut. Control*, **AC-28**, 1067-1073.
- Neuman, C. P. and J. J. Murray (1984). Linearization and sensitivity functions of dynamic robot models. *IEEE Trans. Syst. Man Cybern.*, **SMC-14**, 805-818.
- Polak, E., D. Q. Mayne and Y. Wardi (1983). On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems. *SIAM J. Control and Optimiz.*, **21**, 179-203.
- Sahar, G. S. and J. M. Hollerbach (1986). Planning of minimum-time trajectories for robot arms. *Int. J. Robotics Res.*, **5**, 90-100.
- Shin, K. G. and N. D. McKay (1985). Minimum-time control of robotic manipulators with geometric path constraints. *IEEE Trans. Aut. Control*, **AC-30**, 531-541.
- Stoer, J. and R. Bulirsch (1980). *Introduction to Numerical Analysis*. Springer, Berlin.
- Weinreb, A. and A. E. Bryson (1985). Optimal control of systems with hard control bounds. *IEEE Trans. Aut. Control*, **AC-30**, 1135-1138.

Brief Paper

An Algorithm for the Assignment of System Zeros*

W. A. BERGER,† R. J. PERRY‡ and H. H. SUN§

Key Words—Linear systems; zeros.

Abstract—This paper presents an algorithm for assigning the locations and number of zeros to a linear system through synthesis of the system input and output space couplings. The approach imposes no restrictions on the state-space model, can assign both real and complex zeros using only real arithmetic, and is computationally efficient. Numerical properties of the algorithm are discussed and examples are given to illustrate its performance.

1. Introduction

AN IMPORTANT topic in the synthesis of multivariable systems is the problem of obtaining a system having the desired properties, i.e. the outputs respond in a desirable manner to reference inputs (Fallside, 1977; Patel and Munro, 1982; Ohm *et al.*, 1985). The concepts of eigenvalues and zeros (MacFarlane and Karcnias, 1976; Kouvaritakis and MacFarlane, 1976) form a natural link between frequency-response and state-space approaches, and their locations in the complex plane can be interpreted in terms of geometrical relationships involving the system $\{A, B, C, D\}$.

The eigenvalues and zeros of a system are values that completely specify the system response. The field of eigenvalue assignment by state variable feedback has been well developed, and a number of numerically reliable solutions are available (Patel and Misra, 1984; Kautsky *et al.*, 1985; Miminis and Paige, 1988; Petkov *et al.*, 1986). The approach of Petkov has been proven to be numerically stable (Cox and Moss, 1989). It is assumed here that the system under consideration already has all of its eigenvalues at desired locations.

The zeros represent the nature of the couplings between the system's characteristics modes and its environment. They depend on the matrices B and C and on the way in which these matrices are related to the eigenframework of the matrix A . By suitable variation of the null space of C and the range of B , i.e. suitable choices of input and output couplings, zeros can be placed at any possible locations in the complex plane (MacFarlane and Karcnias, 1976). In terms of system response, the concept of zeros is intimately related to the physical situation in which the system has an identically zero output while the states and inputs are not themselves identically zero.

* Received 17 November 1987; revised 30 August 1988; revised 26 April 1990; revised 3 September 1990; received in final form 26 September 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak.

† Department of Physics/Electronics Engineering, University of Scranton, Scranton, PA 18510, U.S.A.

‡ Department of Electrical Engineering, Villanova University, Villanova, PA 19085, U.S.A.

§ Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, U.S.A. Author to whom all correspondence should be addressed.

During the past decade considerable attention has been paid to the computation of the zeros and especially to the development of reliable numerical software (Laub and Moore, 1978; Emami-Naeini and VanDooren, 1982a, b; Laub, 1985) for this problem. The problem of assignment of system zeros has been addressed in the literature (Kouvaritakis and MacFarlane, 1976; Sadeghi *et al.*, 1986; Bhattacharyya *et al.*, 1986; Perry *et al.*, 1986, 1988), however, these approaches are numerically unstable in general due to their use of ill-conditioned transformations.

This paper considers assigning a prescribed set of element zeros to a linear system described by a state-space model $\{A, B, C, D\}$, where the system internal dynamics A are given and the external input-output couplings $\{B, C, D\}$ are to be synthesized. In particular, we present an algorithm for assigning the values and the number of finite zeros associated with an individual transfer function between any input and output through synthesis of the output space mapping. This new algorithm is a key part of a method for complete assignment of a multi-input multi-output system pole-zero configuration (Berger, 1988), where the problem of zero assignment is first reduced to a sequence of one or more subproblems involving the individual elements of the transfer function matrix.

The assignment problem can be formulated in a similar manner in terms of multivariable transmission zeros and will be presented in another paper. A method was presented (Misra and Patel, 1988) for transmission zero assignment using only feedthrough compensation D , whereas in our approach both the output coupling C and feedthrough term D are being synthesized.

II. Assignment of system zeros

Consider the state-space model of a linear time-invariant system:

$$\begin{aligned}\dot{x} &= Ax + Bu \\ y &= Cx + Du\end{aligned}\quad (1)$$

where x is an n -dimensional vector of system states, u is an m -dimensional vector of system inputs, y is a p -dimensional vector of system outputs, and where λ can be the differential or the delay operator. A , B and C are constant coefficient matrices and D represents some direct "nondynamical" coupling between input and output.

In general, the Transfer Function Matrix $H(\lambda)$ is a p, m -matrix of individual transfer functions in the form:

$$H_{ij}(\lambda) = c_i^T (\lambda I - A)^{-1} b_j + d_{ij} = G_{ij} \frac{\prod_{k=1}^{l_{ij}} (\lambda - z_k^{(ij)})}{\prod_{k=1}^{p_k} (\lambda - p_k)} \quad (2)$$

representing the transfer function between input j and output i , where c_i^T is the i th row of C , and b_j is the j th column of B . This form shows the poles and zeros before any pole-zero cancellation has been taken into account. The system poles, p_k , are the eigenvalues of A . Each transfer function has a

gain factor G_g , and l_g zeros z_k which are the finite zeros of the system matrix, the $(n+1)$, $(n+1)$ pencil:

$$S_g(\lambda) = \begin{bmatrix} \lambda I - A & b_g \\ -c_g^T & d_g \end{bmatrix}. \quad (3)$$

The problem considered here is to choose d_g and the input and output coupling vectors b_g and c_g^T such that the resulting closed-loop transfer function H_g has a prescribed set of finite zeros. Assuming that a suitable value of b_g is given, such that the pair (A, b) is controllable, the following algorithm can be used to determine c_g^T and d_g . By duality, c_g^T could be given, and a dual form of the following algorithm, using (A^T, c_g^T) in place of (A, b) , used to determine the input mapping b_g .

We now present an algorithm which, given A , b , a prescribed gain G and a set of finite zeros, z_k , $k = 1 \dots l$, determines values for c^T and d . The assignment of zeros is achieved in two steps: a Reduction Algorithm and an Assignment Algorithm. The Reduction Algorithm proceeds by reducing the system matrix to a finite structure pencil (Kailath, 1980) of size $(l+1)$, $(l+1)$. Then, the problem is reduced to an eigenvalue assignment problem which is solved using an existing numerically reliable method (Patel and Misra, 1984) based on the implicitly shifted QR algorithm. Finally, the Assignment Algorithm transforms the computed results back to the original coordinate system.

Reduction algorithm. The system matrix (3) is reduced to size $(l+1)$, $(l+1)$ as follows:

Given: The system state matrix A and an input vector b . Compute: The reduction of the system matrix (3) to a finite structure pencil of size $(l+1)$, $(l+1)$, where l is the number of zeros to be assigned.

Step 1. Initialize:

$$F \leftarrow A, \quad g \leftarrow b, \\ \beta \leftarrow 1, \quad U \leftarrow I_n, \quad \eta \leftarrow n.$$

Step 2. Check reduced system size

If $\eta = l$, Stop.

Reduction is completed.

Step 3. Use a Householder transformation Q such that $Q^T F$ is reduced to $\|g\|_2 e_\eta$, where e_η is the last column of I_η .

$$\beta \leftarrow \beta \|g\|_2, \\ \begin{bmatrix} F & g \\ * & * \end{bmatrix} \leftarrow Q^T F Q, \quad U \leftarrow U \begin{bmatrix} Q & 0 \\ 0 & I_{n-\eta} \end{bmatrix}, \\ \eta \leftarrow \eta - 1$$

Goto Step 2.

The similarity transformation using Q does not change the transfer function, but it reduces the effect of g into a scalar $\|g\|_2$ multiplied by a unit vector, which can be accumulated and saved for future use as a gain factor.

At each iteration of Step 3, F and g are reduced in size by 1 and the last row of $Q^T F Q$ is discarded. The cumulative effect of the reduction process is saved in the parameter β and the n , n orthogonal transformation matrix U .

It can be shown that the reduced system, (F, g, v^T, ρ) , where v^T and ρ are as yet unknown, will have the same finite zeros as the system $(A, b, c^T, 0)$. This can be demonstrated by examining the effect of the first iteration of Step 3 on the system matrix (3).

$$\Omega(\lambda) = \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda I - A & b \\ -c^T & 0 \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} \\ = \begin{bmatrix} \lambda I - Q^T A Q & \|b\|_2 e_\eta \\ -c^T Q & 0 \end{bmatrix}. \quad (4)$$

The right-hand side of (4) can then be transformed by unimodular row and column transformations $L(\lambda)$ and $R(\lambda)$ to

$$L(\lambda) \Omega(\lambda) R(\lambda) = \begin{bmatrix} \lambda I - F & g & 0 \\ -v^T & \rho & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (5)$$

Equations (4) and (5) illustrate that the assigned zeros are

invariant to transformations of the reduced system back to the original physical coordinates. For the above Reduction Algorithm, the form of the unimodular matrices $L(\lambda)$ and $R(\lambda)$ can be explicitly derived: $L(\lambda)$ interchanges the last two rows of $\Omega(\lambda)$, and $R(\lambda)$ divides the last column of $\Omega(\lambda)$ by $\|b\|_2$ and reduces the last row to $[0 \dots 0 \ 1]$. It should be clear that the actual algorithm does not use $L(\lambda)$ and $R(\lambda)$, which are not in general orthogonal transformations.

Therefore, by induction, each iteration of Step 3 yields a reduced system with the same finite zeros.

The next step in the generation of the system output mapping is the assignment of the desired finite zeros. The zeros of the reduced system are known to be the eigenvalues of $F - g(v^T/\rho)$, where v^T and ρ are as yet unknown. Reformulating this expression as $F - gk^T$, the problem of assigning zeros is reduced to an eigenvalue assignment problem for which a good numerical method exists (Patel and Misra, 1984). The inputs to this algorithm are the set of desired finite zeros, z_k , $k = 1 \dots l$ (both real and complex-conjugate pairs, arranged consecutively) and the reduced pair (F, g) . On the output, the gain vector k is obtained such that $F - gk^T$ has all the desired eigenvalues, i.e. finite zeros of the system matrix. It is worth to point out that there is no numerical difficulty with the division by ρ in the term v^T/ρ since the term is being determined as a whole. Once k is determined, we can transform the result back to the original coordinate system and find c^T and d using U .

Assignment algorithm.

Given: F , g , β , U from the Reduction Algorithm, a prescribed gain G and desired zeros, z_k , $k = 1 \dots l$.

Compute: k and the values for c and d , such that the system (A, b, c^T, d) has the desired set of zeros.

Case 1, $l = 0$:

$$d = 0, \quad c^T = \frac{G}{\beta} [1 \ 0 \dots 0] U^T$$

Case 2, $l \neq 0$: Use a numerically robust method such as Patel and Misra (1984) to determine k such that the eigenvalues of $F - gk^T$ are equal to the desired zeros, z_k , $k = 1 \dots l$.

Case 2a, $l = n$:

$$d = G, \quad c^T = Gk^T U^T = Gk^T \quad (\text{since } U^T = I_n)$$

Case 2b, $0 < l < n$:

$$d = 0, \quad c^T = \frac{G}{\beta} [k^T \ 1 \ 0 \dots 0] U^T.$$

The parameter β represents the contribution of the vector b in the system gain factor G of (2). The multiplication by the scale factor G/β in the assignment algorithm cancels the contribution of b and produces the desired gain factor G . Note that in cases 1 and 2b the number of trailing zeros involved in the formulas for c^T is equal to $n - l$, i.e. the number of zeros at infinity.

III. Numerical properties of the algorithm

The algorithm is computationally efficient and we have reasons to believe that it has good numerical properties since only orthogonal transformations are used throughout the algorithm, as well as throughout the solution of the eigenvalue assignment problem.

The solution to the assignment problem involves finding an orthogonal matrix N and a state feedback gain row vector k such that $N^T(F - gk^T)N$ is in Real Schur Form with the desired zeros along the diagonal. The algorithm (Patel and Misra, 1984) enables assignment of real as well as complex-conjugate pairs of zeros using only real arithmetic.

The computational work is divided between the reduction algorithm and the assignment algorithm depending on how many zeros are being assigned. For example, if n zeros are being assigned, the reduction step is not performed and the assignment algorithm is applied to a n , n system. At the other extreme, if the number of desired zeros is 0, the assignment algorithm is not performed and the reduction algorithm proceeds through n iterations. At each iteration,

the system size is reduced and thus the size of the transformation matrices and the number of computations performed are reduced. This is a numerically advantageous feature of both our reduction algorithm and the assignment algorithm (Patel and Misra, 1984).

A Householder transformation acting on a s, t matrix requires $2st$ operations (Wilkinson, 1965). Thus, step 3 of our Reduction Algorithm requires 2η operations to compute β and the Householder transformation, $4\eta^2$ operations to update F , and $2\eta n$ operations to update U . The total number of operations required to reduce the system matrix dimension by $(n - l)$ is $O(n^3 - l^3)$, and is equal to

$$n(n+1)[(n+1)+2(2n+1)/3] \\ - l(l+1)[(n+1)+2(2l+1)/3].$$

Computation of the gain vector k requires $O(l^3)$ operations (Patel and Misra, 1984) and execution of our Assignment Algorithm requires $n(l+1)$ operations. The complete reduction and assignment process therefore requires a grand total of $O(n^3)$ operations.

IV. Examples

The algorithms for assignment of zeros were implemented as MATLAB (Moler, 1980) procedures and the computations performed using double precision on a VAX computer. The numerical method (Emami-Naeini and VanDooren, 1982a, b) for computation of system zeros was used to verify the assigned zeros.

Example 1. Consider the rectangular system (Kouvaritakis and MacFarlane, 1976; Emami-Naeini and VanDooren, 1982a, b)

$$A = \begin{bmatrix} -2 & -6 & 3 & -7 & 6 \\ 0 & -5 & 4 & -4 & 8 \\ 0 & 2 & 0 & 2 & -2 \\ 0 & 6 & -3 & 5 & -6 \\ 0 & -2 & 2 & -2 & 5 \end{bmatrix} \quad B = \begin{bmatrix} -2 & 7 \\ -8 & -5 \\ -3 & 0 \\ 1 & 5 \\ -8 & 0 \end{bmatrix} \\ C = \begin{bmatrix} 0 & -1 & 2 & -1 & -1 \\ 1 & 1 & 1 & 0 & -1 \\ 0 & 3 & -2 & 3 & -1 \end{bmatrix} \quad D = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Let the desired zeros of H_{21} be $(1, -2 + j3, -2 - j3, -4)$ and desired gain factor $G_{21} = 2$. Since there are 4 zeros to be assigned, only one reduction of the system matrix is required. After Step 3 of the Reduction Algorithm we have:

$$\beta = 11.9164 \\ \begin{bmatrix} F & R \\ * & * \end{bmatrix} = \begin{bmatrix} -4.6130 & 4.1617 & -2.6873 & 5.0377 & 0.3549 \\ -1.9989 & 0.7758 & -0.8320 & 3.0860 & -0.3164 \\ -4.9295 & 4.9860 & -1.5045 & 4.5547 & 1.3079 \\ -4.3160 & 5.7122 & -2.2332 & 4.0600 & 0.7948 \\ -7.4661 & 7.9749 & -3.9327 & 4.8355 & 4.2817 \end{bmatrix} \\ U = \begin{bmatrix} -0.6713 & -0.6713 & -0.2518 & 0.0839 & -0.1678 \\ -0.3859 & 0.6141 & -0.1447 & 0.0482 & -0.6713 \\ -0.1447 & -0.1447 & 0.9457 & 0.0181 & -0.2518 \\ 0.0482 & 0.0482 & 0.0181 & 0.9940 & 0.0839 \\ 0.6141 & -0.3859 & -0.1447 & 0.0482 & -0.6713 \end{bmatrix}$$

In the Assignment Algorithm we have:

$$k^T = [15.4619 \quad -38.4597 \quad 8.4297 \quad -28.8935] \\ c^T = [1.8000 \quad -5.5167 \quad 1.7667 \quad -4.9667 \quad 3.5333] \\ d = 0.$$

The computed assigned zeros are:

$$0.999999999999986 \\ -1.999999999999995 + j2.999999999999996 \\ -1.999999999999995 - j2.999999999999996 \\ -4.000000000000008$$

Example 2. This example illustrates the numerical accuracy of our algorithm. In the system of Example 1, H_{11} has zeros $(19/9, -1, -2, -3)$, and gain factor $G_{11} = 9$. Let b equal the first column of B and let the desired zeros and gain factor be the same as those of H_{11} . Application of the reduction and assignment algorithms should yield c^1 equal to row 1 of C i.e. c^1 . The expected and computed results are:

$$c_1 \quad c \\ 0 \quad 0.000000000000000 \\ -1 \quad -1.000000000000001 \\ 2 \quad 2.000000000000000 \\ -1 \quad -1.000000000000002 \\ -1 \quad -0.999999999999999$$

Example 3. This example will illustrate the numerical accuracy of the method presented here in contrast to the results obtained using a method based on non-orthogonal similarity transformations and reduction to phase-variable canonical form (Perry *et al.*, 1986).

Consider a 10th order system:

$$A = \text{diagonal}(2^{10}, 2^{-1}, \dots, 2^{-9}), \\ b = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]^T, \\ c^1 = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1], \quad d = 0.$$

Let the desired zeros and gain factor be the same as those of the above system. Application of the reduction and assignment algorithms should yield the above exact c^1 .

The results shown below indicate close agreement between the computed c and the exact c , whereas the third column of the table, showing the equivalent parameter $c^{(*)}$ as computed using the method with non-orthogonal transformations (Perry *et al.*, 1986), has fewer digits of accuracy.

exact c	computed c	$c^{(*)}$
1	1.000000000000001	0.999999638421474
1	1.000000000000001	0.99999268820255
1	1.000000000000000	0.999998504286334
1	1.000000000000000	0.999996861130893
1	0.999999999999999	0.999992895241870
1	1.000000000000000	0.999982764739651
1	0.999999999999999	0.999899446282154
1	1.000000000000001	1.000046929698883
1	1.000000000000003	0.999969502995560
1	0.999999999999997	1.000114188382926

V. Conclusions

An algorithm for assigning a prescribed set of finite zeros associated with the individual transfer functions between any input and output has been presented. The algorithm imposes no restrictions on the state-space model, and enables assignment of both real and complex zeros using only real arithmetic. The method is efficient since reduced size matrices are used in each step, and has favorable numerical behavior since only orthogonal transformations are performed.

Examples were presented to illustrate the numerical performance of the algorithm. Furthermore, a comparison of the results with those obtained from a method using non-orthogonal similarity transformations showed the superior accuracy of the new algorithm presented here.

A MATLAB implementation of this algorithm is available from the authors.

References

- Berger, W. A. (1988). Complete parameter control in multivariable systems. A numerically robust approach with applications. Ph.D. Thesis, Drexel University, PA.
- Bhattacharyya, S. P., J. W. Howze and S. O. Majidi (1986). Zero assignment by measurement feedback. *IEEE Trans. Aut. Control*, AC-31.
- Cox, C. L. and W. F. Moss (1989). Backward error analysis for a pole assignment algorithm. *SIAM J. Matrix Anal. Applic.*, 10, 446-456.

- Emami-Naeini, A. and P. VanDooren (1982a). Computation of zeros of linear multivariable systems. *Automatica*, **18**, 415-430.
- Emami-Naeini, A. and P. VanDooren (1982b). On computation of transmission zeros and transfer functions. *IEEE Symp. Decision and Control*, Orlando, FL.
- Fallside, F. (Ed.) (1977). *Control System Design by Pole-Zero Assignment*. Academic Press, London.
- Kailath, T. (1980). *Linear Systems*. Prentice-Hall, Englewood Cliffs, NJ.
- Kautsky, J., N. K. Nichols and P. VanDooren (1985). Robust pole assignment in linear state feedback. *Int. J. Control*, **41**, 1129-1155.
- Kouvaritakis, B. and A. G. J. MacFarlane (1976). Geometric approach to the analysis and synthesis of system zeros, parts 1 and 2. *Int. J. Control*, **23**, 149.
- Laub, A. J. (1985). Numerical linear algebra aspects of control design computations. *IEEE Trans. Aut. Control*, **AC-30**, 97.
- Laub, A. J. and B. C. Moore (1978). Calculation of transmission zeros using QZ techniques. *Automatica*, **14**, 557-566.
- MacFarlane, A. G. J. and N. Karamias (1976). Poles and zeros of linear multivariable systems: A survey of the algebraic, geometric and complex-variable theory. *Int. J. Control*, **24**, 33-74.
- Mimnis, G. S. and C. C. Paige (1988). A direct algorithm for pole assignment of time-invariant multi-input linear systems using state feedback. *Automatica*, **24**, 343-356.
- Misra, P. and R. V. Patel (1988). Transmission zero assignment in linear multivariable systems. *IEEE Conf. on Decision and Control*, Austin, TX.
- Moler, C. B. (1980). MATLAB User's Guide. Technical Report CS81-1, Department of Computer Science, University of New Mexico, Albuquerque, NM.
- Ohm, D. Y., J. W. Howze and S. P. Bhattacharyya (1985). Structural synthesis of multivariable controllers. *Automatica*, **21**, 35-55.
- Patel, R. V. and N. Munro (1982). *Multivariable System Theory and Design*. Pergamon, Oxford.
- Patel, R. V. and P. Misra (1984). Numerical algorithms for eigenvalue assignment by state feedback. *Proc. IEEE*, **72**.
- Perry, R. J., H. H. Sun and W. A. Berger (1986). Complete pole-zero control in multivariable systems. *Proc. IEEE 1986 ICCAS*, San Jose, CA.
- Perry, R. J., H. H. Sun and W. A. Berger (1987a). Transfer function matrix realization in multivariable systems. *Proc. Third Int. Symp. on Application of Multivariable System Techniques*, Plymouth, UK; also (1988) *Trans. Inst. Meas. Control*, London, UK.
- Perry, R. J., H. H. Sun and W. A. Berger (1987b). Sensitivity of pole-zero configurations in transfer function matrix realization. *Proc. IEEE 1987 ICCAS*, Philadelphia, PA.
- Perry, R. J., H. H. Sun and W. A. Berger (1988). Determination of transfer function matrix in multivariable systems. *IEEE Trans. Aut. Control*, **AC-33**.
- Petkov, P. H. R., N. D. Christov and M. M. Konstantinov (1986). A computational algorithm for pole assignment of linear multi-input systems. *IEEE Trans. Aut. Control*, **AC-31**.
- Sadeghi, T., D. H. Hoitsma and L. Schoenberg (1986). A control law for pole/variant zero placement. *Proc. ACC*, San Francisco, CA.
- Stewart, G. W. (1973). *Introduction to Matrix Computations*. Academic Press, New York.
- Wilkinson, J. H. (1965). *The Algebraic Eigenvalue Problem*. Oxford University Press, London.

A Lyapunov Robustness Bound for Linear Systems with Periodic Uncertainties*

W. L. CHEN† and J. S. GIBSON‡

Key Words—Lyapunov robustness; stability; periodic coefficients; robust control

Abstract—For linear systems involving uncertain parameters with known, constant nominal values and uncertain perturbations that vary sinusoidally with time, Lyapunov robustness analysis is used to determine a stability bound, or margin, for the amplitudes of the parameter perturbations. This bound is the size of a hypercube in parameter space for which asymptotic stability is guaranteed. The bound, which is based on a quadratic Lyapunov function that depends linearly on parameter perturbations, varies with the frequency of the uncertain parameter perturbations. The bound is asymptotically proportional to the square root of this frequency as it becomes large.

1. Introduction

NUMEROUS RECENT papers have used quadratic Lyapunov functions to develop robustness bounds for linear systems with uncertain parameters. Some papers (Hyland and Bernstein, 1987; Patel and Toda, 1980; Yedavalli, 1985; Yedavalli and Liang, 1986; Zhou and Khargonekar, 1987) have dealt with robustness analysis only, while some (Bernstein, 1987; Haddad and Bernstein, 1988; Keel *et al.*, 1988; Petersen, 1988; Petersen and Hollot, 1986; Zhou and Khargonekar, 1988) have used Lyapunov-based robustness analysis as a basis for design of robust controllers. A common feature of the references just cited and most related work is that a single Lyapunov function is used for the entire set of parameters for which stability is guaranteed. Because of this, the Lyapunov robustness analysis applies to time-varying uncertain parameters (although the nominal plant must be constant). However, the robustness bounds, or margins, produced by such analysis involve only the magnitude of parameter variations; the analysis cannot detect how the allowable magnitude of uncertain time-varying parameters depends on their frequency.

In Leal (1988) and Leal and Gibson (1990), a quadratic Lyapunov function was developed that varies linearly with uncertain plant parameters. Because of the linear dependence on parameters, the method used is called a first-order method. For all but one example to date, this first-order method has yielded larger robustness bounds than the sharpest possible method based on parameter-independent Lyapunov functions [see Leal (1988) and Leal and Gibson (1990)]. The first-order method does not apply to problems with time-varying uncertainties, though.

This paper extends the approach in Leal (1988) and Leal and Gibson (1990) to linear systems in which the nominal

system is time-invariant but the perturbations in uncertain coefficients vary sinusoidally with time. As in these studies, the Lyapunov function here varies linearly with uncertain parameters. The first-order term in the Lyapunov matrix satisfies a differential equation in which the forcing term contains the sinusoidal perturbations from the nominal plant. As a result, the Lyapunov function and the resulting robustness margin depend on the frequency of the parameter perturbations. In general, the robustness margin is proportional to the square root of the frequency of the perturbations at large frequencies.

2. The state and Lyapunov equations

We consider the system

$$\dot{x}(t) = A(t)x(t), \quad x(0) = x_0 \quad (2.1)$$

where $x(t)$ is a real n -vector and the $n \times n$ matrix $A(t)$ has the form

$$A(t) = A(t, p) = A_0 + G(p) \sin \omega t \quad (2.2)$$

where the real matrix $G(p)$ is a linear function of the constant parameter vector $p = [p_1, p_2, \dots, p_m]^T \in R^m$ and the real matrix A_0 is constant and independent of p .

Definition 1 A real $n \times n$ matrix function $P(t)$ is a *Lyapunov matrix* for $A(t)$ if (i) $P(t)$ is periodic with period $2\pi/\omega$, (ii) $P(t)$ is symmetric and positive definite for each t , its maximum eigenvalue is bounded uniformly in t and its minimum eigenvalue is bounded away from 0 uniformly in t , (iii) $P(t)$ is piecewise continuously differentiable and the real symmetric matrix

$$Q(t) = -[P(t) + A(t)^T P(t) + P(t)A(t)] \quad (2.3)$$

is non-negative. Furthermore, $P(t)$ is a *strict Lyapunov matrix* for $A(t)$ if $Q(t)$ is positive definite with its minimum eigenvalue bounded away from 0 uniformly in t .

Theorem 1. The system (2.1) is uniformly exponentially stable if and only if there exists a strict Lyapunov matrix for $A(t)$.

We assume that the eigenvalues of A_0 all have negative real parts, so that for each positive definite symmetric real $n \times n$ matrix Q_0 there exists a unique positive definite symmetric real $n \times n$ matrix P_0 satisfying

$$A_0^T P_0 + P_0 A_0 = -Q_0 \quad (2.4)$$

We will factor Q_0 uniquely as

$$Q_0 = LL^T \quad (2.5)$$

where L is a real $n \times n$ lower triangular matrix with positive diagonal elements.

3. The first-order method

We define

$$P(t, p) = P_0 + P_1(t, p) \quad (3.1)$$

where, for each value of the parameter vector p , $P_1(t, p)$ is

* Received 16 June 1989; revised 21 May 1990; received in final form 24 October 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. F. Curtain under the direction of Editor H. Kwakernaak.

† Mechanical, Aerospace and Nuclear Engineering, University of California, Los Angeles 90024, U.S.A.

‡ Author to whom all correspondence should be addressed.

the unique periodic solution to

$$\begin{aligned} \dot{P}_1(t, p) + A_0^T P_1(t, p) + P_1(t, p) A_0 \\ = -[G^T(p)P_0 + P_0 G(p)] \sin \omega t. \end{aligned} \quad (3.2)$$

That (3.2) has exactly one periodic solution follows from the fact that the eigenvalues of A_0 all have negative real parts. The matrix $P_1(t, p)$ is a linear function of the parameter vector p , since $G(p)$ is. Furthermore, $P_1(t, p)$ has the form

$$P_1(t, p) = P_a(p) \cos \omega t + P_b(p) \sin \omega t \quad (3.3)$$

where the real symmetric $n \times n$ matrices $P_a(p)$ and $P_b(p)$ are the unique $n \times n$ matrices that satisfy the equations

$$-\omega P_a(p) + A_0^T P_b(p) + P_b(p) A_0 = -[G^T(p)P_0 + P_0 G(p)] \quad (3.4)$$

$$\omega P_b(p) + A_0^T P_a(p) + P_a(p) A_0 = 0. \quad (3.5)$$

To motivate our terminology, we note that, if the solution $P(t, p)$ to (2.3) for fixed $Q = Q_0$ is expanded as a Taylor series in p , the zero-order term is P_0 and the first-order term is $P_1(t, p)$.

Next, we define some quantities that will be useful in determining whether $P(t, p)$ is a Lyapunov matrix for $A(t, p)$. First,

$$W(t, p) = L^{-1}[G^T(p)P_1(t, p) + P_1(t, p)G(p)]L^{-1} \sin \omega t \quad (3.6)$$

where L is the matrix in (2.5). For a matrix M ,

$$\sigma_{\max}(M) = \text{maximum singular value of } M; \quad (3.7)$$

$$\begin{aligned} \rho(M) = \text{spectral radius of } M \text{ (for square } M) \\ = \max \{|\lambda|, \lambda \text{ is an eigenvalue of } M\}. \end{aligned} \quad (3.8)$$

Since P_0 is a strict Lyapunov matrix for A_0 , the following two conditions, together, are sufficient for $P(t, p)$ to be a strict Lyapunov matrix for $A(t, p)$ for a given p :

Condition 1. $\rho(P_0^{-1}P_1(t, p))$ is bounded strictly below 1 uniformly in t .

Condition 2. $\sigma_{\max}(W(t, p))$ is bounded strictly below 1 uniformly in t .

Definition 2. (Hypercube in R^m). For $s \geq 0$,

$$C(s) = \{p = [p_1 \ p_2 \ \cdots \ p_m]^T : \max_i |p_i| \leq s\}.$$

We note that $C(s)$ is the convex hull of its 2^m vertices.

Definition 3. If f is a real-valued function defined on $C(1)$, then

$$\mu_1(f) = \max \{f(p) : p \text{ is a vertex of } C(1)\}.$$

Lemma 1. Let $\{\xi_1, \xi_2, \dots, \xi_k\}$ be a finite collection of points in a linear space, let S be the convex hull of $\{\xi_1, \xi_2, \dots, \xi_k\}$, and let f be a convex function defined on S . Then $\max \{f(\xi) : \xi \in S\} = f(\xi_j)$ for some j .

The proof of this lemma, given in Leal (1988), is elementary.

We recall that $\sigma_{\max}(\cdot)$ is a norm for any space of finite dimensional matrices. Hence $\sigma_{\max}(\cdot)$ is a convex function on any such space. Also, for a fixed matrix M , $\rho(M^T M N) = \sigma_{\max}(M N M^T)$ is a convex function on the space of symmetric matrices N of a given dimension.

Now we will use the foregoing definitions and facts to estimate the largest hypercube $C(s)$ such that, for each p in the interior of $C(s)$, $P(t, p)$ is a strict Lyapunov matrix for $A(t, p)$. The final result will be Theorem 2.

Since $P_a(p)$ and $P_b(p)$ are linear in p and since a convex function of a linear function is convex, Lemma 1 yields

$$\begin{aligned} \rho(P_0^{-1}P_a(sp)) = s\rho(P_0^{-1}P_a(p)) \leq s\mu_1(\rho(P_0^{-1}P_a)), \\ p \in C(1) \text{ and } s \geq 0, \end{aligned} \quad (3.9)$$

and similarly for $P_b(p)$. Then, since a sum of convex

functions is convex and since the square of a nonnegative convex function is convex,

$$\rho(P_0^{-1}P_1(t, sp)) \leq s\sigma_{1\max}(Q_0), \quad p \in C(1), \quad s, t \geq 0, \quad (3.10)$$

where

$$\sigma_{1\max}(Q_0) = \mu_1([\rho(P_0^{-1}P_a)^2 + \rho(P_0^{-1}P_b)^2]^{1/2}). \quad (3.11)$$

We factor $G(p)$ as

$$G(p) = G_0 G_1(p) \quad (3.12)$$

where G_0 is independent of p and $G_1(p)$ is linear in p . Since $[L^{-1}P_a(p)G_0]$ and $[G_1(p)L^{-1}]$ are linear functions of p , $\sigma_{\max}(L^{-1}P_a(p)G_0)$ and $\sigma_{\max}(G_1(p)L^{-1})$ are convex functions of p , and similarly for $P_b(p)$. Therefore, Lemma 1 and elementary properties of $\sigma_{\max}(\cdot)$ yield

$$\begin{aligned} \sigma_{\max}(W(t, p)) \leq 2\mu_1(\sigma_{\max}(L^{-1}P_1(t, p)G_0 \sin \omega t)) \\ \cdot \mu_1(\sigma_{\max}(G_1 L^{-1})), \quad p \in C(1). \end{aligned} \quad (3.13)$$

[Recall $\mu_1(\cdot)$ from Definition 3]. From

$$\begin{aligned} 2(P_a(p) \cos \omega t + P_b(p) \sin \omega t) \sin \omega t \\ = P_b(p) + P_a(p) \sin 2\omega t - P_b(p) \cos 2\omega t, \end{aligned} \quad (3.14)$$

it follows that

$$\begin{aligned} 2\sigma_{\max}(L^{-1}P_1(t, p)G_0 \sin \omega t) \leq \sigma_{\max}(L^{-1}P_b(p)G_0) \\ + [\sigma_{\max}(L^{-1}P_a(p)G_0)^2 + \sigma_{\max}(L^{-1}P_b(p)G_0)^2]^{1/2}. \end{aligned} \quad (3.15)$$

Hence (3.13), (3.15) and Lemma 1 yield

$$\begin{aligned} \sigma_{\max}(W(t, sp)) \leq s^2 \sigma_{1\max}(Q_0) \sigma_{2\max}(Q_0), \\ p \in C(1), \quad s, t \geq 0, \end{aligned} \quad (3.16)$$

where

$$\begin{aligned} \sigma_{1\max}(Q_0) = \mu_1(\sigma_{\max}(L^{-1}P_b G_0)) \\ + \mu_1([\sigma_{\max}(L^{-1}P_a G_0)^2 + \sigma_{\max}(L^{-1}P_b G_0)^2]^{1/2}) \end{aligned} \quad (3.17)$$

and

$$\sigma_{2\max}(Q_0) = \mu_1(\sigma_{\max}(G_1 L^{-1})). \quad (3.18)$$

Now we define

$$s_1(Q_0) = 1/\max \{\sigma_{1\max}(Q_0), (\sigma_{1\max}(Q_0) \cdot \sigma_{2\max}(Q_0))^{1/2}\} \quad (3.19)$$

For p in the interior of $C(s_1(Q_0))$, it follows from (3.10)–(3.11) that Condition 1 holds and it follows from (3.16)–(3.18) that Condition 2 holds. Therefore, we have the following theorem, which is the main result of the paper.

Theorem 2. Let Q_0 be a positive definite, real symmetric $n \times n$ matrix. For each p in the interior of $C(s_1(Q_0))$, $P(t, p)$ is a strict Lyapunov matrix for $A(t, p)$.

From (3.3)–(3.5), it follows that $P_1(t, p)$ is proportional to $1/\omega$ for large ω . From (3.11) and (3.17) then, it follows that $\sigma_{1\max}(Q_0)$ and $\sigma_{1\max}(Q_0)$ are asymptotically proportional to $1/\omega$. Since $\sigma_{2\max}(Q_0)$ is independent of ω , $(\sigma_{1\max}(Q_0) \cdot \sigma_{2\max}(Q_0))^{1/2}$ dominates $\sigma_{1\max}(Q_0)$ for large ω , so that $s_1(Q_0)$ is proportional to $\omega^{1/2}$ for large ω .

4. Numerical solution of the Lyapunov equations

Eliminating $P_b(p)$ from (3.4) and (3.5) yields

$$\begin{aligned} \omega^2 P_a(p) + A_0^T P_a(p) + P_a(p) A_0^2 + 2A_0^T P_a(p) A_0 \\ = \omega[G^T(p)P_0 + P_0 G(p)]. \end{aligned} \quad (4.1)$$

The algorithm in Bartels and Stewart (1972) for solving standard Lyapunov algebraic equations can be generalized in the following way to solve (4.1) for $P_a(p)$. Let U be a real unitary matrix such that $U^T A_0 U = A_1$, where A_1 has quasi Schur form. Then premultiplying (4.1) by U^T , postmultiplying by U and inserting UU^T where needed yields

$$\begin{aligned} \omega^2 \tilde{P}_a(p) + A_1^T \tilde{P}_a(p) + \tilde{P}_a(p) A_1^2 + 2A_1^T \tilde{P}_a(p) A_1 \\ = \omega U^T [G^T(p)P_0 + P_0 G(p)] U \end{aligned} \quad (4.2)$$

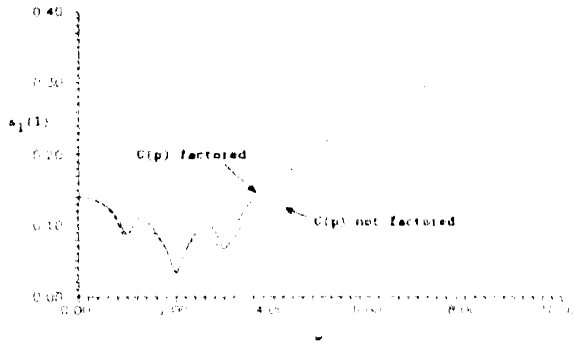


FIG. 1. $s_1(I)$ vs frequency of uncertain parameter perturbations

where $\tilde{P}_a(p) = U^T P_a(p) U$. The various 1×1 , 1×2 , 2×1 and 2×2 blocks of (4.2) can be solved recursively as in Bartels and Stewart (1972).

5 Example

The matrices A_0 and $G(p)$ in (2.2) are the following 4×4 matrices:

$$A_0 = \begin{bmatrix} 0 & I \\ -K_0 & -D \end{bmatrix} \quad G(p) = \begin{bmatrix} 0 & 0 \\ -K_1 & 0 \end{bmatrix} \quad (5.1)$$

where

$$K_0 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \quad D = 0.05 K_0 \quad K_1 = \begin{bmatrix} p_1 & p_2 \\ p_2 & p_1 \end{bmatrix}. \quad (5.2)$$

Figure 1 shows $s_1(I)$ as a function of ω for $G(p)$ factored as in (3.12) with

$$G_0 = \begin{bmatrix} 0 \\ I \end{bmatrix} (4 \times 2), \quad G_1(p) = [-K_1 \quad 0] (2 \times 4), \quad (5.3)$$

and for $G(p)$ not factored (i.e. $G_0 = I$ and $G_1(p) = G(p)$). That $s_1(I)$ is asymptotically proportional to $\omega^{1/2}$, as predicted in Section 3, is clear from Fig. 1.

Perhaps more interesting are the local minima at $\omega = 1, 2, 3$ and 4 . We recall a classical result for the undamped Mathieu equation [see Nayfeh and Mook (1979) or other standard references]: a parametric excitation of frequency twice that of the nominal system makes the solution to the equation unstable. Thus the local minima at $\omega = 2$ and $\omega = 4$ (twice the natural frequencies of the nominal system) might be expected. Furthermore, results in Chapter 5 of Nayfeh and Mook (1979) for a multi-degree-of-freedom Mathieu equation indicate additional instabilities produced by parametric excitation frequencies equal to sums and/or differences of natural frequencies of the nominal system. [For the one-degree-of-freedom damped Mathieu equation obtained by taking A_0 and $G(p)$ to be 2×2 matrices of the forms in (5.1) and (5.2) with $K_0 = 1$ and $K_1 = p_1$, we have obtained an $s_1(I)$ plot similar to Fig. 1 but with only the local minimum at $\omega = 2$.]

6. Conclusions

For linear systems with uncertain parameters that vary with time so that the known nominal system is time-invariant and the uncertain perturbations in the parameters vary sinusoidally with time, the Lyapunov robustness method here yields a stability-robustness margin that depends on the frequency of the uncertain parameter perturbations. This margin is asymptotically proportional to the square root of the frequency of the uncertain parameter perturbations, and, as the example indicates, the robustness margin is sharp enough to identify important resonances.

Numerical optimization of the matrix L should increase $s_1(Q_0)$ in most examples, as in Leal (1988) and Leal and Gibson (1990) for the case of constant uncertain parameters. Also, some estimates in Section 4 perhaps are not the sharpest possible. However, that $s_1(Q_0)$ is asymptotically proportional to the square root of the parameter frequency is inherent in the method.

The methods of this paper can be applied to the case where the perturbation $G(p) \sin \omega t$ in (2.2) is replaced by $\sum_{i=1}^N H_i(p) \sin(\omega_i t + \phi_i)$. In this case, the matrix $P_1(t, p)$ is

replaced by $\sum_{i=1}^N P_i(t, p)$ where each $P_i(t, p)$ has the form on the right side of (3.3) and satisfies a Lyapunov equation similar to (3.2). Estimates corresponding to those developed in (3.9)–(3.19) are obtained by the same methods but with more uses of the triangle inequality, so that the estimate of the largest hypercube for which stability robustness is guaranteed is cruder than for the case treated in this paper.

Acknowledgement—This research was supported by the United States Air Force under AFOSR Grant 870373.

References

- Bernstein, D. S. (1987) Robust static and dynamic output feedback stabilization: deterministic and stochastic perspectives. *IEEE Trans. Aut. Control*, **AC-32**, 1076–1084.
- Bartels, R. H. and G. W. Stewart (1972). Solution of the equation $AX + XB = C$. *Comm. Assoc. Comp. Mach.*, **15**, 820–826.
- Haddad, W. M. and D. S. Bernstein (1988). Robust reduced-order modeling via the optimal projection equations with Petersen–Hollot bounds. *IEEE Trans. Aut. Control*, **AC-33**, 692–695.
- Hyland, D. C. and D. S. Bernstein (1987). The majorant Lyapunov equation: a nonnegative matrix equation for robust stability and performance of large scale systems. *IEEE Trans. Aut. Control*, **AC-32**, 1005–1013.
- Keel, L. H., S. P. Bhattacharyya and J. W. Howze (1988). Robust control with structured perturbations. *IEEE Trans. Aut. Control*, **AC-33**, 68–78.
- Leal, M. A. (1988). Objective and constraint functions for the analysis and design of robust control systems. PhD dissertation, UCLA, CA.
- Leal, M. A. and J. S. Gibson (1990). Lyapunov robustness bounds for linear systems with uncertain parameters. *IEEE Trans. Aut. Control*, **AC-35**, 1068–1070.
- Nayfeh, A. H. and D. T. Mook (1979). *Nonlinear Oscillations*. Wiley, New York.
- Patel, R. V. and M. Toda (1980). Quantitative measures of robustness for linear state space models. *Proc. Joint Aut. Control Conf.*, Paper TP8-A, San Francisco, CA.
- Petersen, I. R. (1988). Stabilization of an uncertain linear system in which uncertain parameters enter into the input matrix. *SIAM J. Control Optimiz.*, **26**, 1257–1264.
- Petersen, I. R. and C. V. Hollot (1986). A Riccati equation approach to the stabilization of uncertain linear systems. *Automatica*, **22**, 397–411.
- Yedavalli, R. K. (1985). Improved measures of stability robustness for linear state space models. *IEEE Trans. Aut. Control*, **AC-30**, 557–579.
- Yedavalli, R. K. and Z. Liang (1986). Reduced conservatism in stability robustness bounds by state transformation. *IEEE Trans. Aut. Control*, **AC-31**, 863–866.
- Zhou, K. and P. P. Khargonekar (1987). Stability robustness bounds for linear state-space models with structured uncertainty. *IEEE Trans. Aut. Control*, **AC-32**, 621–623.
- Zhou, K. and P. P. Khargonekar (1988). On the stabilization of uncertain linear systems via bound invariant Lyapunov functions. *SIAM J. Control, Optimiz.*, **26**, 1265–1273.

Brief Paper

A Class of Invariant Regulators for the Discrete-time Linear Constrained Regulation Problem*

JEAN-CLAUDE HENNET†‡ and JEAN-PAUL BEZIAT†

Key Words—Discrete-time linear systems; Linear Constrained Regulation Problem; stability; positive invariance; variable regulator; linear programming.

Abstract—Stable dynamic systems admit positively invariant domains associated to their Lyapunov functions. Conversely, some domains can be made positively invariant for systems with state feedback controllers designed in such a way that some associated non-negative definite functions are bound to decrease. In particular, this approach can be used to establish conditions on the gain matrix for Linear Constrained Regulation Problems (LCRP). We construct fixed and variable regulators easy to compute through linear programming, for a class of constrained linear systems.

Introduction

TECHNICAL CONTROL limitations have long been considered a major problem in control engineering. Actually, most control schemes do not integrate constraints in their design. In practice, control laws often have to be complemented by adequate control limiting devices. However, saturated linear controllers may fail to stabilize unstable linear systems.

For some sets of initial conditions, stabilizing saturated linear controllers can be designed by the method of Gutman and Hagander (1985). Their approach rests on the construction of an elliptic positively invariant domain included in the polyhedral domain of constraints and including the set of initial states. The shape of the invariant domains is directly induced by the selected Lyapunov functions. So, the choice of classical quadratic Lyapunov functions is not the most efficient for the Linear Constrained Regulation Problem (LCRP); it does not maximize the size of the domain of initial states for which a stabilizing constrained regulator can be computed. This limitation can be overcome by the use of non-quadratic Lyapunov functions of the type introduced by Rosenbrock (1963). Along this line, some authors (Vassilaki *et al.* 1988; Benzaouia and Burgat, 1988) have proposed methods for constructing polyhedral positively invariant sets better fitted or even perfectly matching the domain of linear constraints.

In particular, for a linear state feedback, the set of control constraints generates a convex polyhedron in the state space. This polyhedron can be made invariant by specific matrix conditions. A simplified version of the invariance conditions

is proposed in this paper. It allows for an easy on-line implementation of the control scheme with possible extensions to adaptive cases (Béziat and Hennet, 1988). This method is based on linear programming. In the feasible cases, it generates a rapidly converging control law; computation of the gain matrix can also be frequently updated to accelerate the convergence speed by taking into account the current state of the system. Global stability of the variable regulating scheme is proven under a local stability condition.

Positive invariance of polyhedral sets

Consider the discrete-time linear dynamical system described by the equation:

$$X_{k+1} = A_k X_k \quad (1)$$

with $X_k \in \mathbb{R}^n$ for any $k \in \mathbb{N}$, $A_k \in \mathbb{R}^{n \times n}$.

Let $R(G, \omega)$ be a not-empty convex polyhedral set defined by:

$$R(G, \omega) = \{X \in \mathbb{R}^n \mid GX \leq \omega\} \quad (2)$$

$G \in \mathbb{R}^{g \times n}$ and ω is a vector in \mathbb{R}^g . The inequalities between vectors are componentwise. For instance, in definition (2), $GX \leq \omega$ stands for $(GX)_i \leq (\omega)_i$, for $i = 1, \dots, g$.

According to the selected pair (G, ω) , $R(G, \omega)$ can be any type of polyhedral set (bounded or unbounded, including or not the origin point). The case of proper cones is also included in this representation, for $\omega = 0$.

By definition, $R(G, \omega)$ is said to be positively invariant for system (1) if and only if:

$$X_k \in R(G, \omega) \Rightarrow X_{k+1} = A_k^+ X_k \in R(G, \omega) \\ \forall k \in \mathbb{N}, \quad \forall k \in \mathbb{N}$$

This definition of positive invariance can be found, in particular, in Lasalle (1976).

The property of Ω -invariance defined in Gutman and Cwikel (1986)—is closely related to this definition. But it combines the positive invariance of a domain of the state space with linear constraints on the control vector and with an asymptotic stability requirement. In contrast, positive invariance of an unbounded polyhedral set does not generally require or imply the asymptotic stability of the state trajectories emanating from this set.

Existence of positively invariant polyhedral sets for system (1) is a generic property which covers different types of dynamical behaviours. However, if $\text{rank } G = n$ and $(\omega)_i > 0$ for $i = 1, \dots, g$, then, positive invariance of $R(G, \omega)$ implies the stability of system (1). This last property can easily be shown as in Bitons (1988a) by selecting as a Lyapunov function of the system:

$$V(X) = \max \left\{ \frac{|(GX)_i|}{(\omega)_i} \right\} \quad (3)$$

The following proposition is valid for any type of

* Received 19 July 1989; revised 3 September 1990; received in final form 28 September 1990. The original version of this paper was presented at the 11th IFAC World Congress Symposium on Automatic Control at the Service of Mankind which was held in Tallinn, Estonia, during August 1990. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor R. V. Patel under the direction of Editor H. Kwakernaak.

† Laboratoire d'Automatique et d'Analyse des Systèmes, 7, avenue du Colonel Roche, 31077 Toulouse, France.

‡ Author to whom all correspondence should be addressed.

polyhedral set $R(G, \omega)$. It provides a necessary and sufficient condition for $R(G, \omega)$ to be a positively invariant set of system (1). This basic result on polyhedral invariant sets was initially established by Bitsoris (1988b) under some more restrictive conditions (in particular, all the components of vector ω were supposed strictly positive). Here, these conditions are relaxed and a more direct proof is presented.

Proposition 1. The convex polyhedral set $R(G, \omega)$ is positively invariant for system (1) if and only if there exists a matrix $H \in \mathcal{R}^{n \times n}$ such that:

$$H_{ij} \geq 0, \text{ for } i = 1, \dots, g; j = 1, \dots, g \tag{4}$$

$$HG = GA_0 \tag{5}$$

$$H\omega \leq \omega \tag{6}$$

Proof. A necessary and sufficient condition for $R(G, \omega)$ to be positively invariant for system (1) is:

$$GA_0X \leq \omega, \quad \forall X \in \mathcal{R}^n; GX \leq \omega. \tag{7}$$

Condition (7) is a special case of inclusion of a polyhedral convex set in an other polyhedral convex set. The extended Farkas' lemma (Hennet, 1989) provides an algebraic characterization of such an inclusion.

Any point of $R(G, \omega)$ also satisfies the set of linear inequalities $P \cdot X \leq \psi$ with $P \in \mathcal{R}^{p \times n}$ and $\psi \in \mathcal{R}^p$ if and only if there exists a (dual) matrix H of $\mathcal{R}^{n \times p}$ with non-negative coefficients satisfying conditions $HG = P$ and $H\omega \leq \psi$. This result can easily be proven by concatenation of necessary and sufficient conditions related to each row P_i of matrix P . By the standard Farkas' lemma (see e.g. Schrijver, 1986), a necessary and sufficient condition for:

$$P_i X \leq \psi_i, \quad \forall X : GX \leq \omega$$

is

$$\exists H_i \in \mathcal{R}^{1 \times n}, H_i G = P_i, \quad H_i \cdot \omega \leq \psi_i, \quad (H_i)_j \geq 0, \quad \forall j = 1, \dots, g$$

The simultaneous satisfaction of all these elementary conditions for any point of $R(G, \omega)$ is equivalent to the existence of p row-vectors H_i satisfying the same types of condition as above. The extended Farkas' lemma is simply obtained by constructing matrix H from these p row-vectors. In particular, by setting $P = GA$ and $\psi = \omega$, the necessary and sufficient condition for $R(G, \omega)$ to be positively invariant can be written:

$$\exists H \in \mathcal{R}^{n \times n} : HG = GA_0, \quad H\omega \leq \omega, \\ H_{ij} \geq 0 \text{ for } i = 1, \dots, g; j = 1, \dots, g. \quad \square$$

Invariance and stability under input constraints

Now, the considered multivariable linear systems can be represented by state-space equations of the following type

$$X_{k+1} = AX_k + BU_k \tag{8}$$

$X_k \in \mathcal{R}^n$ being the state vector, $U_k \in \mathcal{R}^m$ the control vector, $A \in \mathcal{R}^{n \times n}$, $B \in \mathcal{R}^{n \times m}$. The control vector is subject to constraints:

$$-Um \leq U_k \leq UM \tag{9}$$

with $(Um)_i \geq 0, (UM)_i \geq 0$, for $i = 1, \dots, m$ and $k = 0, 1, 2, \dots$. Assuming that the state of the system is observable, we want to control its closed-loop dynamics. Then, the selected control law is a linear state feedback, with $F \in \mathcal{R}^{m \times n}$:

$$U_k = FX_k \tag{10}$$

The closed-loop evolution of the system is described by

$$X_{k+1} = A_0X_k \text{ with } A_0 = A + BF \tag{11}$$

The general problem called LCRP (Linear Constrained Regulation Problem) consists of determining a matrix F such that the state vector of system (11) converges to 0 while respecting constraints (9) and relations (8), (10).

From Kalman and Bertram (1960), it is well-known that if system (11) has all its eigenvalues in the unit circle of the complex plane, it admits elliptic positively invariant domains associated to its quadratic Lyapunov functions. The square root of a quadratic Lyapunov function is a contracting norm for the state vector.

From the properties of equivalence between norms (see e.g. Glazman and Liubitch, 1974), existence of a L_2 contracting norm is equivalent to the existence of a polyhedral norm for which the state is contracting. This norm can be written as in relation (3). Existence of such a norm is equivalent to the positive invariance of $R(G, \omega)$. Hennet and Lasserre (1990) have presented a scheme for constructing a polyhedral positively invariant domain for any asymptotically stable system. The converse problem, analyzed in this paper, is to shape the dynamical behaviour of the controlled system so that a particular domain is made positively invariant. The convex polyhedron generated in the state space by constraints (9) is:

$$R[F, Um, UM] = \{X \in \mathcal{R}^n : -Um \leq FX \leq UM\}.$$

A specialized version of invariance conditions (4-6) is provided by the following proposition

Proposition 2. A sufficient condition for $R[F, Um, UM]$ to be positively invariant is the existence of a pair of non-negative matrices ($H' \in \mathcal{R}^{m \times m}, H \in \mathcal{R}^{m \times m}$) such that:

$$F(A + BF) \leq HF \tag{12}$$

$$\tilde{H}\rho \leq \rho \tag{13}$$

with

$$H = H' + H'', \quad \tilde{H} = \begin{pmatrix} H' & H'' \\ H & H' \end{pmatrix}, \quad \tilde{H} \in \mathcal{R}^{2m \times 2m}$$

and

$$\rho = \begin{pmatrix} UM \\ Um \end{pmatrix}, \quad \rho \in \mathcal{R}^{2m}.$$

This condition is also necessary when rank $F = m$.

Proof

Sufficiency

Invariance of $R[F, Um, UM]$ is readily derived from relations (12) and (13) by a direct application of Proposition 1 to system (11) with:

$$G = \begin{bmatrix} +F \\ -F \end{bmatrix} \text{ and } \omega = \rho$$

Necessity when rank $F = m$

Assume now that $R(G, \omega)$ is a positively invariant set of system (11), with G and ω defined as above

Then, from Proposition 1, there exists a matrix with non-negative elements,

$$\tilde{H} = \begin{pmatrix} \tilde{H}_{11} & \tilde{H}_{12} \\ \tilde{H}_{21} & \tilde{H}_{22} \end{pmatrix},$$

$\tilde{H}_{IJ} \in \mathcal{R}_+^{m \times m}$ for $I, J \in (1, 2)$ such that:

$$G(A + BF) = \tilde{H}G, \quad \tilde{H}\omega \leq \omega.$$

Then, $(\tilde{H}_{11} - \tilde{H}_{12})F = (\tilde{H}_{22} - \tilde{H}_{21})F = F(A + BF)$. If rank $F = m$, this relation implies: $\tilde{H}_{11} - \tilde{H}_{12} = \tilde{H}_{22} - \tilde{H}_{21}$. Set $H = \tilde{H}_{11} - \tilde{H}_{12}$. Let H' be the matrix of the non-negative components of H and H'' the matrix of the non-negative components of $(-\tilde{H})$. Then, matrix $H = H' - H''$ satisfies condition (12). And

$$\begin{cases} (H')_{ij} \leq \min \{(\tilde{H}_{11})_{ij}, \tilde{H}_{22}\}_{ij} \\ (H'')_{ij} \leq \min \{(\tilde{H}_{12})_{ij}, \tilde{H}_{21}\}_{ij} \end{cases} \text{ for } i, j \in (1, \dots, m).$$

Thus, matrices H' and H'' satisfy the necessary condition (13):

$$\begin{pmatrix} H' & H'' \\ H & H' \end{pmatrix} \omega \leq \begin{pmatrix} \tilde{H}_{11} & \tilde{H}_{12} \\ \tilde{H}_{21} & \tilde{H}_{22} \end{pmatrix} \omega \leq \omega. \quad \square$$

Proposition 2 can also be derived from Proposition 4 in Bitsonis (1988b), which treats the case of given non-symmetrical polyhedral domains of the state-space.

The algorithm of the next section solves relations (12), (13) by an indirect technique. A solution matrix H computed by this technique also guarantees the invariance of a given polyhedral domain of the state-space, $R[G, Um, UM]$ with G full-rank. Under this scheme, the class of matrices H can be replaced by the class of matrices \hat{H} with no loss of generality.

In the special case of symmetrical constraints ($Um = UM \neq 0_m$), relation (13) can be equivalently replaced by $(1 - |H|)Um \geq 0$. A necessary condition for this inequality to be satisfied is $(1 - |H|)$ to be an M -matrix (Benzouia and Burgat, 1988).

The LCRP can be solved whenever it is possible to find (H^*, H^*, F) such that:

- Matrix $(A + BF)$ has all its eigenvalues located in the unit circle of the complex plane.
- Positive invariance conditions (12), (13) are verified
- The initial state vector, $X_0 \in R[F, Um, UM]$.

The algebraic formulation of this last condition is:

$$-Um \leq FX_0 \leq UM \quad (14)$$

In this paper, we propose a method for finding an easily computable pair (H, F) belonging to a subclass of solutions of equation (12). This design technique is based on the following proposition:

Proposition 3 If $R[G, Um, UM]$ is a positively invariant domain of system (11), and H an associated solution of the system:

$$G(A + BF) = HG, \quad \hat{H}\rho \geq \rho,$$

then any polyhedral domain $R[Q, Um, UM]$ is also a positively invariant domain of (11) if

$$\begin{cases} Q = \Delta G; \quad \Delta \in \mathcal{R}^{m \times m} \\ \Delta H = H\Delta \end{cases}$$

Proof. $G(A + BF) = HG$ and $\Delta H = H\Delta$ imply $Q(A + BF) = HQ$. Therefore, under the assumed condition $\hat{H}\rho \geq \rho$, $R[Q, Um, UM]$ is also positively invariant. \square

Now, a controller $U_k = FX_k$ letting $R[F, Um, UM]$ positively invariant can be constructed as follows:

- Select a fixed matrix $G \in \mathcal{R}^{m \times m}$ for which $R[G, Um, UM]$ can be made positively invariant by state feedback. Then, $\exists (H \in \mathcal{R}^{m \times m}, F \in \mathcal{R}^{m \times n})$; $HG = G(A + BF)$, $\hat{H}\rho \geq \rho$
- Solve the linear system:

$$\begin{cases} HG = G(A + BDG) \\ HD = DH, \quad D \in \mathcal{R}^{m \times m} \\ \hat{H}\rho \geq \rho \end{cases}$$

- Set $F = DG$.

The subclass of controls obtained by this scheme is characterized by the additional property of letting $R[G, Um, UM]$ positively invariant. It is therefore important to handily select matrix G so as to obtain a nonempty subclass of controls for most feasible problems.

A basic condition for the existence of non-negative vectors (Um, UM) such that $R[G, Um, UM]$ can be made positively invariant by state feedback is that $\text{Ker}(G)$ should be an (A, B) -invariant subspace in the sense of Wonham (Hennet and De Bona Castelan Neto, 1990).

For any system in its minimal representation, the assumption $\text{rank } B = m$ with $m \leq n$ is quite general. Under this assumption, there exist matrices $B^X \in \mathcal{R}^{m \times n}$ such that

$$B^X B = I_{m \times m} \quad (15)$$

In particular, we can select the left pseudo-inverse of B ; $B^X = (B^T B)^{-1} B^T$. The kernel of B^X is the quotient space $\mathcal{R}^n / \text{Im}(B)$, which is an (A, B) -invariant subspace, since $\mathcal{R}^n / \text{Im}(B) + \text{Im}(B) = \mathcal{R}^n$ (Wonham, 1985).

If we assume $G = B^X$ and apply Proposition 3, we can obtain a subclass of positively invariant controllers by

imposing the following set of conditions:

$$B^X A B H = H B^X A B \quad (16)$$

$$F = H B^X - B^X A \quad (17)$$

$$\hat{H}\rho \leq \rho \quad (18)$$

Equation (16) is obtained from condition $DH = HD$ with $D = H - B^X A B$, since H commutes with itself. And clearly, the choice of F from relation (17) satisfies condition $B^X(A + BF) = H B^X A$. In the case $m \leq n$, a positively invariant controller satisfying conditions (16), (17) and (18) does not generally guarantee the overall stability of the closed-loop system. Some additional stability conditions have to be introduced to get a stabilizing positively invariant regulator.

An algorithm for solving the LCRP

A possible way to simplify the stability analysis is to impose as a positively invariant set a polytope $R(G, w)$ with $R(G, w) \subset R[F, Um, UM]$, $w > 0$ and $\text{rank } G = n$ (Vassilaki et al., 1988). Under this assumption, the function $V(X)$ defined by relation (3) is positive definite and can be chosen as a candidate Lyapunov function for system (11). In this paper, the polytope to be maintained invariant will be directly constructed by completing $R[F, Um, UM]$ under the assumption: $(Um)_i > 0$, $(UM)_i > 0$, for $i = 1, \dots, m$.

In the case $m < n$, it is always possible to add dummy control variables $(U_k)_{m+1}, \dots, (U_k)_n$ to add $n - m$ independent columns in B so that $\text{rank } B = n$, and to impose the following constraints:

$$-(Um)_{m+i} \leq (U_k)_{m+i} \leq (UM)_{m+i} \quad (19)$$

with, for instance, $(Um)_{m+i} = (UM)_{m+i} = \theta$, for $i = 1, \dots, n - m$, and $\theta > 0$ as small as desired.

Under these extra conditions, it can be assumed that X_k and U_k have the same dimension, n , and that $\text{rank } B = n$. In this case, the resolution of the LCRP can be directly obtained from the following Proposition:

Proposition 4 If $m = n$ and $\text{rank } B = n$, conditions (20), (21) and (22) guarantee the positive invariance of $R[F, Um, UM]$ and the asymptotic stability of the controlled system (11)

$$B^{-1} A B H = H B^{-1} A B \quad (20)$$

$$F = H B^{-1} - B^{-1} A \quad (21)$$

$$\hat{H}\rho \leq \rho \quad (22)$$

Proof. From relations (20) and (21), we can derive condition (12):

$$\begin{aligned} F(A + BF) &= (H B^{-1} - B^{-1} A) B H B^{-1} \\ &= H(H B^{-1} - B^{-1} A) \\ &= H F \end{aligned}$$

Thus, under inequality (22), $R[F, Um, UM]$ is a positively set of the closed-loop system.

Under condition (21), the dynamic matrix of the closed-loop system is:

$$A_0 = B H B^{-1} \quad (23)$$

Matrices A_0 and H being similar, asymptotic stability of the controlled system (11) is equivalent to the asymptotic convergence to 0 of the control sequence, which satisfies, for $k = 0, 1, \dots$, the recurrent relation (24)

$$\begin{aligned} U_{k+1} &= F(A + BF)X_k \\ &= H F X_k \\ U_{k+1} &= H U_k \end{aligned} \quad (24)$$

Constraint (22) consists of the two inequalities:

$$H^* U M + H^* U m < U M$$

$$H^* U M + H^* U m < U m$$

The summing up of these two inequalities yields:

$$(H^* + H^*)(U M + U m) < U M + U m$$

Therefore, matrix $|H| = (|H_{ij}|)$ and vector $W = UM + Um$ satisfy:

$$|H|W \leq (H^+ + H^-)W < W.$$

W is a positive vector of \mathcal{M}^n . Then $(I - |H|)$ is an M -Matrix, and from a classical result presented in Lasalle (1976), it is a necessary and sufficient condition for system (24) to be asymptotically stable. \square

The coefficients of matrices H^+ and H^- can be taken as the unknown variables of a linear programming problem, denoted problem (II) and formulated as follows:

$$\min \epsilon \quad (25)$$

subject to:

$$\tilde{H}\rho \leq \epsilon\rho \quad (26)$$

$$B^{-1}AB(H^+ - H^-) = (H^+ - H^-)B^{-1}AB \quad (27)$$

$$B^{-1}AX_0 - Um \leq (H^+ - H^-)B^{-1}X_0 \leq B^{-1}X_0 + UM \quad (28)$$

Note that inequalities (28) simply express that the initial state should belong to $R[F, Um, UM]$ (relation 14) for a gain matrix F satisfying relation (21).

An efficient way of solving the LCRP can be derived from the following proposition.

Proposition 5. If problem (II) has a solution (H^+, H^-, ϵ) with $\epsilon < 1$, then the LCRP is solved by using the control:

$$U_k = F \cdot X_k \quad \text{with} \quad F = (H^+ - H^-)B^{-1} - B^{-1}A. \quad (29)$$

Proof. From Proposition 4, positive invariance of $R[F, Um, UM]$ and asymptotic stability of the closed-loop system derive from the respect of conditions (27) and (26) when $\epsilon < 1$; and condition (28) in problem (II) guarantees:

$$X_0 \in R[F, Um, UM]. \quad \square$$

If the optimal value of ϵ , ϵ^* , is strictly smaller than 1, compute F by (29) and set $U_k = FX_k$.

Note that the solution of problem (II) explicitly depends on the initial state of the system, X_0 , through inequalities (28). However, it is clear that the computed gain matrix F can also stabilize the system from any other initial point belonging to $R[F, Um, UM]$, and that the control trajectory always remains feasible. If the initial state of the system is not perfectly known, a design technique imposing the invariance of a domain containing the domain of possible initial states (Gutman and Hagander, 1985; Vassilaki *et al.*, 1988) is probably more appropriate.

If $\epsilon^* \geq 1$, stability of the closed-loop system is not achieved by this algorithm.

The fact of imposing the invariance of $R[B^{-1}, Um, UM]$ and relations (20) and (21) constrains the closed-loop eigenvectors to belong to some subspaces. But under the assumptions of this paragraph, this is not a severe restriction. Any set of n independent directions can generate an invariant domain of the closed-loop system.

It is only the size of the domain of stabilizable initial states which may be reduced by using a control belonging to the investigated subclass. This possibility constitutes the only case of "conservativeness" of the proposed algorithm, when a suitable feedback exists but cannot be found by solving problem (II).

The efficiency of this algorithm can be improved when the current state of the system can be observed. A variable regulating scheme can then be implemented. The gain matrix is periodically updated by solving a linear problem denoted (Π_k) , similar to problem (II) except for relations (28), in which the initial state vector, X_0 , is replaced by the current state vector, X_k :

$$\min \epsilon_k \quad (30)$$

subject to:

$$\tilde{H}_k \cdot \rho \leq \epsilon_k \cdot \rho \quad (31)$$

$$B^{-1}AB(H_k^+ - H_k^-) = (H_k^+ - H_k^-)B^{-1}AB \quad (32)$$

$$B^{-1}AX_k - Um \leq (H_k^+ - H_k^-)B^{-1}X_k \leq B^{-1}AX_k + UM. \quad (33)$$

The variable regulating law is:

$$U_k = F_k X_k \quad (34)$$

and F_k is computed from the optimal solution $(H_k^+, H_k^-, \epsilon_k)$ of problem (Π_k) by relation:

$$F_k = (H_k^+ - H_k^-)B^{-1} - B^{-1}A. \quad (35)$$

The closed-loop evolution of the system is described by:

$$X_{k+1} = (A + BF_k)X_k.$$

Replace F_k by its expression (35). It yields

$$X_{k+1} = BH_k B^{-1}X_k. \quad (36)$$

The eigenvalues of H_k are also the eigenvalues of $BH_k B^{-1}$.

Proposition 6. If there exists an integer r such that problem (Π_r) has the optimal solution (H_r, ϵ_r) with $\epsilon_r < 1$, then the optimal solution of problem (Π_k) exists and verifies $\epsilon_k < 1$ for any integer $k \geq r$, and the controlled system is asymptotically stable.

Proof. From Proposition 5, relations (20), (21) and (22) are satisfied whenever the optimal solution of problem (Π_k) is such that $\epsilon_k < 1$. Then, from Proposition 4, $X_{k+1} \in R[F_k, Um, UM]$, and (H_k, ϵ_k) is also a feasible solution of problem (Π_{k+1}) . Consequently, the set of feasible solutions of problem (Π_{k+1}) is not empty and by the choice of the objective function, we must have: $\epsilon_{k+1} \leq \epsilon_k$, and, by induction, $\epsilon_r < 1$ implies $\epsilon_k < 1$ for any integer $k \geq r$.

Note that in the case of time-varying linear systems, local stability conditions $\epsilon_k < 1$ for $k \geq r$ do not automatically imply the global stability of the system. Asymptotic stability of the system under the variable feedback law can be proven by showing the existence of a Lyapunov function for the closed-loop system. Since vector W has positive components, the function $L(X)$, defined as follows, is positive definite:

$$L(X) = \max_i \left\{ \frac{|(B^{-1}X)_i|}{W_i} \right\} \quad (37)$$

Constraints (31) consists of the two inequalities:

$$H_k^+ UM + H_k^- Um \leq \epsilon_k UM$$

$$H_k^+ UM + H_k^- Um \leq \epsilon_k Um$$

The summing up of these two inequalities yields:

$$|H_k|W \leq \epsilon_k W \quad (38)$$

The difference between two successive values of function $L(\cdot)$ is $\Delta L_k = L(X_{k+1}) - L(X_k)$. From relation (36), derive $B^{-1}X_{k+1} = H_k B^{-1}X_k$. Then, majorate $L(X_{k+1})$ using (38):

$$\begin{aligned} L(X_{k+1}) &= \max_i \left\{ \frac{|(B^{-1}X_{k+1})_i|}{W_i} \right\} \\ &= \max_i \left\{ \frac{|(H_k B^{-1}X_k)_i|}{W_i} \right\} \\ &= \frac{1}{W_i} \left| \sum_{j=1}^n (H_k)_{ij} \frac{W_j}{W_i} (B^{-1}X_k)_j \right| \\ &\leq \frac{(|H_k|W)_i}{W_i} \max_j \frac{|(B^{-1}X_k)_j|}{W_j} \\ &\leq \epsilon_k \cdot \max_i \frac{|(B^{-1}X_k)_i|}{W_i}. \end{aligned}$$

Then, $L(X_{k+1}) \leq \epsilon_k L(X_k)$ and since $\epsilon_k < 1$ for $k \geq r$, $\Delta L_k < 0$. $L(X)$ is a Lyapunov function of the sequence of state vectors, which asymptotically converges to 0: $\lim_{k \rightarrow \infty} X_k = 0$. \square

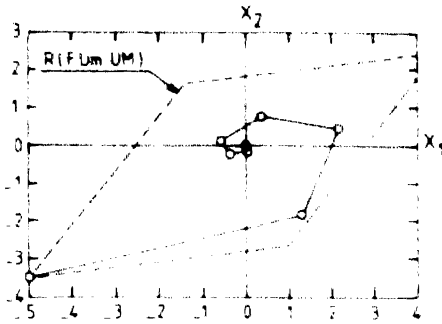


FIG. 1 Trajectory of the state-vector X_k with the constant regulator.

Example

Consider the second order system described by the state-space equation:

$$X_{k+1} = \begin{pmatrix} 1.7 & -3.3 \\ 1.3 & 0.3 \end{pmatrix} X_k + \begin{pmatrix} 3.0 & 2.0 \\ -2.0 & 2.0 \end{pmatrix} U_k$$

The control vector is subject to the following constraints:

$$- \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix} \leq U_k \leq \begin{pmatrix} 1.0 \\ 1.4 \end{pmatrix},$$

and the initial state vector is

$$X_0 = \begin{pmatrix} -5.0 \\ -3.5 \end{pmatrix}$$

The unforced system is unstable. The eigenvalues of matrix A are: $\lambda_{1,2} = 1 \pm j1.95$.

The constant regulator gives the optimal values

$$\epsilon^* = 0.93, \quad H = \begin{pmatrix} 0.38 & -0.35 \\ 0.93 & 0.0 \end{pmatrix}$$

and the control law $U_k = FX_k$ is:

$$U_k = \begin{pmatrix} -0.076 & 0.536 \\ -0.544 & 0.384 \end{pmatrix} X_k$$

Simulation results are presented in Fig. 1.

The same problem can be solved using the variable regulator. The simulation results presented in Fig. 2 show that the variable regulator considerably increases the speed of convergence.

The same system now has to be controlled from a different initial state vector

$$X_0 = \begin{pmatrix} 3.5 \\ 3.0 \end{pmatrix}$$

The successive values of the rate of convergence ϵ_k obtained by the linear programming algorithm are: $\epsilon_0 = 1.53$, $\epsilon_1 = 1.68$, $\epsilon_2 = 0.96$, $\epsilon_3 = 0.92$, $\epsilon_4 = 0.82$, $\epsilon_5 = 0.6$.

We can see in Fig. 3 that although $\epsilon_0 > 1$, the system converges since we get $\epsilon_k < 1$ after two steps. The origin can

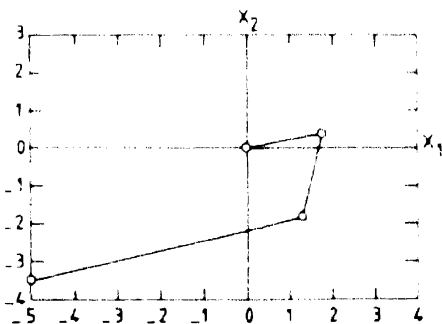


FIG. 2 Trajectory of the state-vector X_k with the variable regulator.

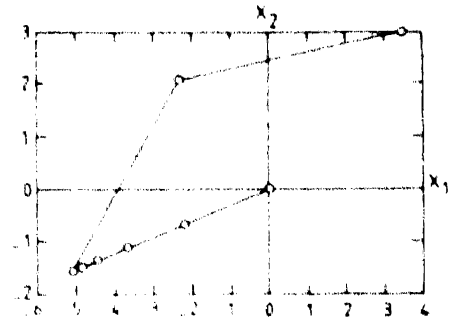


FIG. 3 Trajectory of the state-vector X_k with the variable regulator in a case $\epsilon_0 > 1$.

be reached in 3 time units. So, in particular, X_0 belongs to the maximal Ω -invariant set defined and constructed in Cwikel and Gutman (1986) and Gutman and Cwikel (1987).

Since $\epsilon_0 > 1$, the constant regulator is unable to stabilize the system. In contrast, the variable regulating scheme can be satisfactorily applied.

Conclusions

Many linear constrained regulation problems can be solved by constructing positively invariant domains associated with Lyapunov functions. The invariance conditions obtained by this approach can be used as constraints on the gain matrix of the control law. A linear simplified version of these conditions is presented in this paper. The LCRP can then be solved by a standard linear program. The selected objective function to be maximized is denoted ϵ_0 . It measures the rate of convergence of the system to the origin.

Two types of regulators are proposed. In the fixed scheme, the same gain matrix is applied at each period, while in the variable scheme, a new gain matrix is computed at each period, from the knowledge of the current state of the system.

If $\epsilon_0 < 1$, stability of the closed-loop system is guaranteed with the fixed and with the variable controller, but the variable regulator may considerably increase the speed of convergence.

If $\epsilon_0 > 1$, the fixed regulator is unable to stabilize the system. The variable regulator generates a feasible control as long as the solution of the linear program remains numerically finite. Several iterations of the algorithm can then be computed off-line. And stability of the closed-loop system can often be obtained by getting $\epsilon_k < 1$ after some periods of time. But if the constraints are too severe, the process should rather be stopped to avoid divergence. This variable control law can also be very efficient in an adaptive context, with *a-priori* unknown parameters. Then, at each updating time k , the best current estimates of matrices A and B are updated by a recursive identification algorithm and the computation of the gain matrix also directly uses the information on the current state of the system through resolution of problem (Π_k) .

Acknowledgements—The authors want to thank Professor G. Bitsoris and the anonymous referees for their helpful comments on a previous version of the paper, which was also presented at the XIth IFAC World Congress in Tallinn.

References

- Benzaouia, A. and Ch. Burgat (1988). The regulator problem for a class of linear systems with constrained control. *Syst. Control Lett.*, **10**, 357–363.
- Béziat, J.-P. and J.-C. Hennet (1988). Stability and invariance conditions in generalized predictive control. *IMACS Int. Symp. on System Modelling and Simulation*, Cetraro, Italy, pp. 163–167.
- Bitsoris, G. (1988a). Positively invariant polyhedral sets of discrete-time linear systems. *Int. J. Control*, **47**, 1713–1726.
- Bitsoris, G. (1988b). On the positive invariance of polyhedral sets for discrete-time systems. *Syst. Control Lett.*, **11**, 243–248.

- Cwikel, M. and P. O. Gutman (1986). Convergence of an algorithm to find maximal state constraint sets for discrete-time linear dynamical systems with bounded control and states. *IEEE Trans. Aut. Control*, **AC-31**, 457-459.
- Glazman, I. and Y. Liubitch (1974). *Analyse Linéaire dans les Espaces de Dimension Finie*. Editions MIR, Moscow.
- Gutman, P. O. and P. Hagander (1985). A new design of constrained controllers for linear systems. *IEEE Trans. Aut. Control*, **AC-30**, 22-23.
- Gutman, P. O. and M. Cwikel (1986). Admissible sets and feedback control for discrete-time linear systems with bounded control and states. *IEEE Trans. Aut. Control*, **AC-31**, 373-376.
- Gutman, P. O. and M. Cwikel (1987). An algorithm to find maximal state constraint sets for discrete-time linear dynamical systems with bounded control and states. *IEEE Trans. Aut. Control*, **AC-32**, 251-254.
- Hennet, J. C. (1989). Une extension du lemme de Farkas et son application au problème de régulation linéaire sous contraintes. *C. R. Acad. Sciences*, t. 308, série I, pp. 415-419.
- Hennet, J. C. and E. De Bona Castelan Neto (1990). Invariance and stability by state feedback for constrained linear systems. Note LAAS-CNRS (submitted to the European Control Conference, Grenoble).
- Hennet, J. C. and J. B. Lasserre (1990). Spectral characterization of linear systems admitting positively invariant polytopes. Note LAAS 90010 (submitted to *Math. Control Signals and Systems*).
- Kalman, R. E. and J. E. Bertram (1960). Control systems analysis and design via the second method of Lyapunov. *Trans. A.S.M.E.*, **D82**, 394-400.
- Lasalle, J. P. (1976). *The stability of Dynamical Systems*. SIAM Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.
- Rosenbrock, H. N. (1963). A method of investigating stability. *Proc. IFAC*, 590-594.
- Schrijver, A. (1986). *Theory of Linear and Integer Programming*. Wiley, Chichester, U.K.
- Vassilaki, M., J. C. Hennet and G. Bitsoris (1988). Feedback control of linear discrete-time systems under state and control constraints. *Int. J. Control*, **47**, 1727-1735.
- Wonham, W. M. (1985). *Linear Multivariable Control—A Geometric Approach*. Springer, Berlin.

Continuous-time LQ Regulator Design by Polynomial Equations*

A. CASAVOLA,† M. J. GRIMBLE,‡ E. MOSCA,†§ and P. NISTRI†

Key Words—Optimal control; linear systems; multivariable systems; polynomials.

Abstract—The deterministic continuous-time LQ output regulation problem is solved by polynomial equations as an alternative to the usual Riccati equation approach. In particular, it is shown that Riccati-based and polynomial methods are fully conceptually equivalent in steady-state LQ regulation.

1. Introduction

THIS PAPER deals with the classic Linear Quadratic Output Regulation (LQOR) problem. Only the continuous-time case is considered here; the discrete-time case having been recently reported elsewhere (Mosca and Nistri, 1989).

A continuous-time, linear, time-invariant state-space representation of the plant to be output regulated is considered

$$\begin{cases} \dot{x}(t) = \Phi x(t) + Gu(t) \\ y(t) = Hx(t) \end{cases} \quad (1)$$

with $x(t) \in \mathcal{R}^n$, $u(t) \in \mathcal{R}^m$ and $y(t) \in \mathcal{R}^p$. The problem is to find, if it exists, an input variable $u(\cdot) \in L_2$ minimizing the quadratic-cost

$$J = \int_0^\infty (\|y(t)\|_{\Psi_y}^2 + \|u(t)\|_{\Psi_u}^2) dt \quad (2)$$

for any initial state $x(0)$. In (2) $\Psi_y = \Psi_y^T > 0$, $\Psi_u = \Psi_u^T > 0$, $\|v(t)\|_{\Psi_y}^2 = v^T(t)\Psi_y v(t)$ and the prime denotes transpose. By Parseval's Lemma, the above cost can be expressed in terms of the Laplace transforms of $y(t)$ and $u(t)$

$$\frac{1}{2\pi j} \int_{-\infty}^{\infty} (\|y(s)\|_{\Psi_y}^2 + \|u(s)\|_{\Psi_u}^2) ds \quad (3)$$

It is well known that under stabilizability and detectability assumptions on the triplet (Φ, G, H) , problem (1)–(3) can be solved in state-feedback form by using the unique nonnegative definite solution of the relevant algebraic Riccati equation. Moreover, the resulting closed-loop system turns out to be asymptotically stable.

The aim of this paper is to provide a direct matrix-fraction

approach to the problem. In this way, the solution is obtained by, first, solving a spectral factorization problem; and, next, finding the minimum-degree solution with respect to a “dummy” polynomial matrix of a pair of bilateral Diophantine equations. The specific “minimum-degree” property that identifies the required solution will be made precise in Section 2.

The problem was previously addressed by Kučera (1983) for scalar input plants and completely reachable pairs (Φ, G) . Because of these rather restrictive assumptions, the solution can be obtained in terms of a single Diophantine equation (Kučera, 1983). In Section 2 it is shown that, in the general case, two bilateral Diophantine equations must be solved simultaneously. Another related contribution is Grimble (1987), where the polynomial solution to the LQ stochastic regulator with complete state information was given for the discrete-time case.

One of the reasons for considering a polynomial solution for the standard deterministic LQOR problem is to show that Riccati-based and polynomial methods are fully conceptually equivalent, as far as steady-state (semi-infinite horizon) results are concerned. In particular, stabilizability and/or detectability requirements in the Riccati equation approach are replaced by conditions on the stability of greatest common left and right divisors of polynomial matrices.

The reader is referred to Kučera (1979), whose notation is adopted hereafter as much as possible considering the differences between the discrete and continuous-time cases. For any real rational matrix $R(s)$, $R^*(s) := R'(-s)$. Further, for any polynomial matrix $P(s)$ in the indeterminate s the following notations or definitions are assumed hereafter: $\partial P(s)$ denotes the degree of $P(s)$; $P(s)$ is said to be *row-reduced* if the matrix of the coefficients of the highest power of s in each row of $P(s)$ has full row-rank; similarly, $P(s)$ is said to be *column-reduced* if the matrix of the coefficients of the highest power of s in each column of $P(s)$ has full column-rank; a square polynomial matrix $P(s) = P_0 s^n + P_1 s^{n-1} + \dots + P_n$ is said to be *regular* if its leading matrix coefficient P_0 is nonsingular. Any regular polynomial matrix is both row-reduced and column-reduced. The opposite implication is in general false.

2. Main results

As is well known, the problem (1)–(3) only depends on a completely observable subsystem of (1) obtainable via Kalman's canonical decomposition. Thus, from the outset assume that

(A.1) (Φ, H) is a completely observable pair.

Hereafter all quantities are assumed to be Laplace transforms. If $y(s)$ denotes the output of (1) due to $x(0)$ and the input signal is denoted by $u(s)$,

$$y(s) = HA^{-1}(s)[x(0) + Bu(s)] \quad (4)$$

where $A(s)$ and B are the following polynomial matrices

$$A(s) := sI - \Phi \quad (5)$$

$$B := G. \quad (6)$$

* Received 23 June 1989; revised 3 January 1990; revised 18 June 1990; received in final form 14 September 1990. The original version of this paper was presented at the IFAC Workshop on System Structure and Control which was held in Prague, Czechoslovakia during September 1989. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor V. Kučera under the direction of Editor H. Kwakernaak.

† Dipartimento di Sistemi ed Informatica, Università di Firenze, Via S. Marta 3-50139 Firenze, Italy.

‡ Industrial Control Unit, University of Strathclyde, 50 George Street, Glasgow G1 1QE, U.K.

§ Author to whom all correspondence should be addressed.

Hereafter, for the sake of simplicity, the argument s will be omitted unless it is required to avoid possible confusion. The quadratic cost (3) can be conveniently rewritten as

$$J = \frac{1}{2\pi j} \int_{\mu}^{\infty} (y^* \Psi_y y + u^* \Psi_u u) ds \quad (7)$$

Let us also rewrite (4) as follows

$$y(s) = HA^{-1}(s)x(0) + HB_2(s)A_2^{-1}(s)u(s) \quad (8)$$

where $B_2(s)A_2^{-1}(s)$ is a right-coprime (rc) column-reduced matrix-fraction description of $A^{-1}B$. Hence, $\partial_{ii}B_2 < \partial_{ii}A_2$, $i = 1, 2, \dots, m$, where $\partial_{ii}B_2$ denotes the column-degree of the i th columns of B_2 (Kailath, 1980).

Substituting (8) in (7), obtain

$$J = \frac{1}{2\pi j} \int_{\mu}^{\infty} [u^* A_2^* (A_2^* \Psi_u A_2 + B_2^* \Psi_y B_2) A_2^{-1} u + u^* A_2^* B_2^* \Psi_u A^{-1} x(0) + x'(0) A^{-*} \Psi_y B_2 A_2^{-1} u + x'(0) A^{-*} \Psi_u A^{-1} x(0)] ds \quad (9)$$

where $A_2^* := (A_2^{-1})^*$ and $\Psi_u := H^* \Psi_y H$. Let $E(s)$ be a Hurwitz polynomial matrix solving the following spectral factorization problem

$$E^* E = A_2^* \Psi_u A_2 + B_2^* \Psi_y B_2 \quad (10)$$

Since A_2 is column-reduced, $B_2 A_2^{-1}$ strictly proper, $\Psi_u > 0$, from Kučera (1980) it follows that E is column-reduced. By using (10) and adding and subtracting $x'(0) A^{-*} \Psi_y B_2 E^{-1} E^{-*} \Psi_u A^{-1} x(0)$ in (9), obtain

$$J = J_1 + J_2 \quad (11)$$

with

$$J_1 := \frac{1}{2\pi j} \int_{\mu}^{\infty} L^* L ds$$

$$J_2 := \frac{1}{2\pi j} \int_{\mu}^{\infty} x'(0) A^{-*} (I - \Psi_y B_2 E^{-1} E^{-*} B_2^*) \Psi_u A^{-1} x(0) ds$$

$$L := E^{-*} B_2^* \Psi_u A^{-1} x(0) + E A_2^{-1} u \quad (12)$$

Note that J_2 does not depend on u . In order to simplify the above expression for J_1 , let us consider the following bilateral Diophantine equation

$$E^* Y + Z A = B_2^* \Psi_u \quad (13)$$

It is temporarily assumed that a solution (Y, Z) of (12) exists with $\partial_{ii}Z < \partial_{ii}E^*$, $i = 1, 2, \dots, m$, where $\partial_{ii}Z$ denotes the row-degree of the i th row of Z . Under this assumption, (12) becomes

$$L = Yx(s) + (E - YB_2)A_2^{-1}u + E^{-*}Zx(0) \quad (14)$$

Further, let us temporarily assume that the following polynomial equation can be jointly solved along with (13)

$$XA_2 + YB_2 = E \quad (15)$$

Substitution of (15) into (14) gives

$$L = (Yx + Xu) + E^{-*}Zx(0) \quad (16)$$

Consequently the cost index J_1 can be split into the following three components

$$J_1 = J_3 + J_4 + J_5 \quad (17)$$

$$J_3 := \frac{1}{2\pi j} \int_{\mu}^{\infty} x'(0) Z^* E^{-1} E^{-*} Z x(0) ds \quad (18)$$

$$J_4 := \frac{1}{2\pi j} \int_{\mu}^{\infty} \{ (Yx + Xu)^* E^{-*} Z x(0) + x'(0) Z^* E^{-1} (Yx + Xu) \} ds \quad (19)$$

$$J_5 := \frac{1}{2\pi j} \int_{\mu}^{\infty} (Yx + Xu)^* (Yx + Xu) ds \quad (20)$$

where J_5 does not depend on u . Further, if X and Y are constant matrices, i.e. $\partial X = \partial Y = 0$ (Lemma 2 below verifies

that this property holds true), by using the Residue Theorem, it is easy to prove that J_4 is identically zero. This follows from the fact that $\partial_{ii}Z^* < \partial_{ii}E$, and hence $Z^* E^{-1}$ is a strictly proper stable transfer-matrix, and that $u(\cdot)$ and $x(\cdot)$ are in L_2 . Thus, minimization of J amounts to minimizing J_3 .

Finally, premultiplying both sides of (15) by E^* and taking into account (10), one gets

$$E^* X - ZB = A_2^* \Psi_u \quad (21)$$

The following lemma summarizes the above discussion.

Lemma 1 Provided that

(i) (13) and (21) [or (13) and (16)] admit a solution (X, Y, Z) with $\partial_{ii}Z < \partial_{ii}E^*$, $i = 1, 2, \dots, m$, $\partial X = \partial Y = 0$ and X nonsingular; and

(ii) $J_2 + J_1$ is bounded;

the solution of the LQOR problem is given by

$$u(s) = -X^{-1}Yx(s) \quad (22)$$

with X and Y specified in (i), and correspondingly,

$$J_{\min} = J_2 + J_1 \quad \square$$

A condition under which (i) of Lemma 1 is fulfilled is given by next lemma whose proof is given in the Appendix

Lemma 2. Let the greatest common left divisors (GCLDs) of A and B in (5) and (6) be strictly Hurwitz. Then, there is a unique solution (X, Y, Z) of (13) and (21) [or (13) and (15)] such that

$$\partial_{ii}Z < \partial_{ii}E^*, \quad i = 1, 2, \dots, m \quad (23)$$

$\partial X = \partial Y = 0$, and X nonsingular. \square

The unique solution (X, Y, Z) referred to in Lemma 2, will be called the *minimum row-degree solution* w.r.t. Z .

Boundedness of $J_2 + J_1$ is clearly guaranteed by the stability of the closed-loop system. This, in turn, if the plant has no unstable hidden modes, is fulfilled if and only if E is strictly Hurwitz. In fact, $\det E$ is proportional to the characteristic polynomial of the closed loop system made up by the plant $B_2 A_2^{-1}$ together with the control law (22). The next lemma, whose proof is reported in the Appendix, gives a sufficient condition for E to be strictly Hurwitz

Lemma 3. If (A.1) holds then the spectral factor E is strictly Hurwitz. \square

The previous lemmas show that the LQ output regulation problem can be solved provided that (Φ, H) is a completely observable pair and the GCLD's of A and B are stable. This is the same as assuming that the given plant (Φ, G, H) , which in general need not be completely observable, has all its observable-unreachable eigenvalues stable. The results are summarized in the following theorem.

Theorem 1. Consider the LQ output regulation problem (1)–(3) for the plant $\Sigma = (\Phi, G, H)$. Then,

1. The problem is solvable if and only if the GCLDs of A_0 and B_0 are stable, where $A_0 := sI - \Phi_0$ and $B_0 := G_0$ and $\Sigma_0 = (\Phi_0, G_0, H_0)$ is a completely observable subsystem of Σ obtainable via Kalman's canonical observability decomposition of Σ .
2. Provided that the solvability condition is fulfilled, the optimal input signal is given by (22), where X and Y are obtained by first solving the spectral factorization problem (10) and next finding the *minimum row-degree solution* w.r.t. Z of the pair of bilateral Diophantine equations (13) and (21) [or (13) and (15)], viz. $\partial_{ii}Z < \partial_{ii}E^*$.
3. The overall closed-loop system is asymptotically stable if and only if the plant Σ has no unstable hidden modes. \square

It is interesting to explore the connection between the polynomial solution (X, Y, Z) given in Theorem 1 (part 2) with the classic Riccati-based solution of the LQOR problem. This is considered in Theorem 2, whose assertion can be easily proved, by showing that the triplet (25) fulfills (13) and (21) [or (13) and (15)] and satisfies the row-degree inequalities $\partial_{ii}Z < \partial_{ii}E^*$, $i = 1, 2, \dots, m$.

Theorem 2. Let the plant $\Sigma = (\Phi, G, H)$ be stabilizable and detectable. Also let P denote the unique symmetric non-negative definite solution of the matrix algebraic Riccati equation

$$P\Phi + \Phi^*P - PG^* \Psi_u^{-1}GP + \Psi_v = 0 \quad (24)$$

Then, the minimum row-degree solution w.r.t. Z of the Diophantine equations (13) and (21) [or (13) and (15)] is given by

$$X^*X = \Psi_u, \quad X^*Y = G^*P, \quad Z = B_2^*P \quad \square \quad (25)$$

3. Discussion and examples

The polynomial solution to the LQOR problem given in Mosca and Nistri (1989) looks like a direct translation into the discrete-time context of the one covered in the present paper. In fact, in both cases the solution—if it exists—is given in terms of a spectral factorization problem and a pair of bilateral Diophantine equations. Nevertheless, the degree constraints for the discrete-time case turn out to be different from (23). The discrete-time case degree constraints, in the present continuous-time context, would translate as follows

$$\partial Z < \partial E^*. \quad (26)$$

This condition is weaker than the row-degree inequality (23). If E is *regular*—a property that is fulfilled in the standard discrete-time case discussed in Mosca and Nistri (1989)—(23) is equivalent to (26). The next example shows that in the continuous-time case, where E need not be regular (26) has to be replaced by (23) in order to get the desired constant solution (X, Y) of (13) and (21).

Example 1. Let

$$\Phi = \begin{bmatrix} -1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\Psi_v = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Psi_u = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A possible choice for A_2 and B_2 is

$$A_2 = \begin{bmatrix} s+1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Since $B_2^* \Psi_u = 0_{2,2}$ and A_2 is Hurwitz, one can set $E = A_2$. Consequently, the solution in Theorem 1 (part 2) is as follows

$$X = I_2, \quad Z = Y = 0_{2,2}.$$

On the contrary, (26) does not yield either a unique solution of (13) and (21) or necessarily $\partial X = \partial Y = 0$. In fact, it can be checked that

$$X = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & 0 \\ s+1 & s+\frac{1}{2} \end{bmatrix}, \quad Z = \begin{bmatrix} 0 & 0 \\ -1 & -1 \end{bmatrix}$$

is one possible solution of (13) and (21) fulfilling (26). \square

There are in general, *two* bilateral Diophantine equations, viz. (13) and (21) [or, equivalently (13) and (15)], that must be solved with the row-degree constraints (23) in order to finding the LQ optimal feedback-gain matrix $F = X^{-1}Y$. In particular, the reader is referred to Mosca *et al.* (1990) and Hunt *et al.* (1987) where the need to solve two Diophantine equations is thoroughly investigated. Nevertheless an example is now presented to aid our understanding of this problem.

Example 2. Let

$$\Phi = \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$H = [1 \quad 1], \quad \Psi_u = 1, \quad \Psi_v = 1.$$

Notice that (Φ, G) is not completely reachable, whereas (Φ, H) is completely observable. We find

$$A = \begin{bmatrix} s-1 & 0 \\ 0 & s+\frac{1}{2} \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

A GCLD of A and B is

$$\begin{bmatrix} 1 & 0 \\ 0 & s+\frac{1}{2} \end{bmatrix}$$

which is stable. Thus, according to Lemma 2 the problem is solvable, and, according to Theorem 1 (part 3), the resulting LQ optimal feedback stabilizes the plant being the only plant hidden eigenvalue $\lambda = -\frac{1}{2}$. We also find

$$A_2 = s-1, \quad B_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

and, via spectral factorization, $E = s + \sqrt{2}$. Further, $A_2^* = -s-1$; $B_2^* = [1 \quad 0]$, $E^* = -s + \sqrt{2}$, which implies $\partial Z = 0$.

(13) and (21) [or (13) and (15)], give

$$X = 1, \quad Y = \begin{bmatrix} v_1 = \sqrt{2} + 1, & v_2 = \frac{2(2\sqrt{2}-1)}{7} \end{bmatrix},$$

$$Z = \begin{bmatrix} z_1 = v_1, & z_2 = \frac{2(2\sqrt{2}-1)}{7} \end{bmatrix}$$

Hence,

$$F = - \begin{bmatrix} \sqrt{2} + 1 & \frac{2(2\sqrt{2}-1)}{7} \end{bmatrix}$$

and

$$\Phi + GF = \begin{bmatrix} -\sqrt{2} & \frac{2(2\sqrt{2}-1)}{7} \\ 0 & -\frac{1}{2} \end{bmatrix}.$$

(15) alone yields $X = 1$, $v_1 = \sqrt{2} + 1$, $z_1 = \sqrt{2} + 1$. However, it does not provide any information on v_2 and z_2 . \square

Conclusions

The deterministic continuous-time LQ output regulation problem has been solved via spectral factorization and a pair of bilateral Diophantine equations. Stabilizability and/or detectability requirements in the Riccati equation approach were replaced by conditions on the stability of greatest common left and right divisors of polynomial matrices.

Acknowledgement—This work was partially supported by Italian MURST and CNR.

References

- Casavola, A., E. Mosca and P. Nistri (1989). Deterministic LQ regulator design by polynomial equations. *IFAC Workshop on System Structure and Control*, Prague, 53–56.
- Grimble, M. J. (1987). Relationship between polynomial and state-space solutions for the optimal regulator problem. *Syst. Control Lett.*, **8**, 411–416.
- Hunt, K. J., M. Sebek and M. Grimble (1987). Optimal multivariable LQG control using a single Diophantine equation. *Int. J. Control*, **46**, 1445–1453.
- Kailath, T. (1980). *Linear Systems*. Prentice-Hall, Englewood Cliffs, N.J.
- Kučera, V. (1979). *Discrete Linear Control*. Wiley, New York.
- Kučera, V. (1980). Stochastic multivariable control: A polynomial equation approach. *IEEE Trans. Aut. Control*, **AC-25**, 913–919.
- Kučera, V. (1983). Linear quadratic control, state space vs polynomial equations. *Kybernetika*, **19**, 185–195.
- Mosca, E. and P. Nistri (1989). A direct polynomial approach to LQ regulation. *Lecture Notes of the Workshop on the Riccati Equation in Control, Systems, and Signals*. Pitagora Editrice, Bologna, pp. 8–11.
- Mosca, E., L. Giarè and A. Casavola (1990). On the polynomial equations for the MIMO LQ stochastic regulator. *IEEE Trans. Aut. Control*, **AC-35**, 320–322.

Appendix

Proof of Lemma 2. It is known (Kucera, 1979) that (13) and (21) are solvable provided that the GCLDs of A and B are strictly Hurwitz. All the solutions of (13) and (21) are given by:

$$X \approx X_0 - TB; \quad Y \approx Y_0 + TA; \quad Z \approx Z_0 - E^*T \quad (27)$$

where (X_0, Y_0, Z_0) is a solution of (13) and (21) and T is any polynomial matrix of compatible dimensions. By the Division Theorem for Polynomial Matrices (Kailath, 1980) there exists a unique pair (Q, R) such that

$$Z_0 = E^*Q + R, \quad \partial_n R < \partial_n E^*.$$

Substituting this expression in (27), one obtains $Z \approx R + E^*(Q - T)$. Therefore the minimum row-degree solution of (13) and (21) w.r.t. Z is obtained by taking $T = Q$ in (27):

$$X \approx X_0 - QB; \quad Y \approx Y_0 + QA; \quad Z \approx R. \quad (28)$$

It remains to be shown that $\partial X \approx \partial Y \approx 0$. This can be concluded by first noting that from (10)

$$\partial_n E = \partial_n A_2. \quad (29)$$

This can be seen by considering that since $\partial_n H_2 < \partial_n A_2$ and $\Psi_u > 0$,

$$2\partial_n E = \partial(E^*E)_n = \partial(A_2^*\Psi_u A_2)_n = 2\partial_n A_2$$

where $(E^*E)_n$ denotes the n th diagonal entry of E^*E . From (13), it follows that

$$\partial_n(E^*Y) \leq \partial_n E^*$$

which, in turn, by nonsingularity of E , implies that $\partial Y \approx 0$.

Similarly, from (21) it follows that

$$\partial_n(E^*X) = \partial_n E^*$$

and again, arguing as above, one finds that $\partial X \approx 0$. Finally, by using (23), (29) and the fact that $\Psi_u > 0$, from (21) obtain

$$\partial_n(E^*X) = \partial_n A_2^*.$$

Hence, since E^* and A_2^* are row-reduced, it follows that X is nonsingular.

In Mosca *et al.* (1990) it is shown that the solution of (13) and (21) coincide with that of (13) and (15).

Proof of Lemma 3. First, it will be shown that complete observability of (Φ, G) implies that HB_2 and A_2 are right coprime (rc). In order to prove this, we begin by noting that by PBH test (Kailath, 1980), complete observability of (Φ, H) is equivalent to H and A rc. This, in turn, implies that H and A_1 are rc, if $A^{-1}B = A_1^{-1}B_1$ with A_1 and B_1 left coprime (lc). In fact, $A = \Delta A_1$ and $B = \Delta B_1$ with Δ a GCLD of A and B . Hence, for some polynomial matrices U and V ,

$$I = UA + VH = (U\Delta)A_1 + VH$$

We finally show that if H and A_1 are rc, then HB_2 and A_2 are rc. In fact, consider the transfer-matrices

$$HA_1^{-1}B_1 = HB_2A_2^{-1}$$

for which $\partial \det A_1 = \partial \det A_2$. The expression on the LHS can be minimally realized in state-space form with a state of dimension equal to $\partial \det A_1$ since H and A_1 are rc. Thus, $HB_2A_2^{-1}$ must be also realized in minimal form having the same state dimension. Hence, we conclude that HB_2 and A_2 are rc.

Having proved that HB_2 and A_2 are rc, we next show that $\det E(j\omega) \neq 0, \forall \omega \in \mathcal{R}$ and hence that E is strictly Hurwitz. In order to prove this, we note that HB_2 and A_2 rc implies that, for $\varphi_v = q_v' = \Psi_v^{1/2}$ and $\varphi_u = q_u' = \Psi_u^{1/2}$, $\hat{B}_2 := \varphi_v HB_2$ and $\hat{A}_2 := \varphi_u A_2$, by nonsingularity of q_v and q_u , are rc. In fact for some polynomial matrices \hat{U} and \hat{V}

$$I = UA_2 + VHB_2 = (\hat{U}\varphi_u^{-1})\hat{A}_2 + (\hat{V}\varphi_v^{-1})\hat{B}_2.$$

We now prove that $\det E(j\omega) \neq 0, \forall \omega \in \mathcal{R}$ by contradiction. Suppose there exists $u \in \mathcal{R}^m, u \neq 0$ and $\omega \in \mathcal{R}$ such that

$$0 = \|E(j\omega)u\|^2 = \|\hat{A}_2(j\omega)u\|^2 + \|\hat{B}_2(j\omega)u\|^2$$

This implies that $\hat{A}_2(j\omega)u = \hat{B}_2(j\omega)u = 0$. But since \hat{B}_2 and \hat{A}_2 are rc, for some polynomial matrices U and V , $(\hat{U}\hat{A}_2 + \hat{V}\hat{B}_2)u = Iu \neq 0$. This contradicts singularity of $E(j\omega)$. \square

Guaranteed Properties of Gain Scheduled Control for Linear Parameter-varying Plants*

JEFF S. SHAMMA† and MICHAEL ATHANS‡

Key Words—Gain scheduling; parameter-varying systems; robustness; time-varying systems.

Abstract—Gain scheduling has proven to be a successful design methodology in many engineering applications. However in the absence of a sound theoretical analysis, these designs come with no guarantees on the robustness, performance, or even nominal stability of the overall gain scheduled design.

This paper presents such an analysis for one type of gain scheduled system, namely, a linear parameter-varying plant scheduling on its exogenous parameters. Conditions are given which guarantee that the stability, robustness, and performance properties of the fixed operating point designs carry over to the global gain scheduled design. These conditions confirm and formalize popular notions regarding gain scheduled design, such as the scheduling variable should "vary slowly."

1. Introduction

1.1. *Problem statement.* Gain scheduling (see e.g. Stein, 1980) is a popular engineering method used to design controllers for systems with widely varying nonlinear and/or parameter dependent dynamics, i.e. systems for which a single linear time-invariant model is insufficient. The idea is to select several operating points which cover the range of the plant's dynamics. Then, at each of these points, the designer makes a linear time-invariant approximation to the plant and designs a linear compensator for each linearized plant. In between operating points, the parameters (i.e. "gains") of the compensators are then interpolated, or "scheduled," thus resulting in a global feedback compensator.

Despite the lack of a sound theoretical analysis, gain scheduling is a design methodology which is known to work in a myriad of operating control systems (e.g. jet engines, submarines, and aircraft). However, in the absence of such an analysis, these designs come with no guarantees. More precisely, even though the local point designs may have excellent feedback properties, the global gain scheduled design need not have any of these properties, even nominal stability. In other words, one typically cannot assess *a priori* the guaranteed stability, robustness and performance properties of gain scheduled designs. Rather, any such properties are inferred from extensive computer simulations.

* Received 10 February 1989; revised 14 November 1989; revised 18 June 1990; received in final form 22 August 1990. The original version of this paper was presented at the IFAC Symposium on Nonlinear Control System Design which was held in Capri, Italy during June, 1989. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor R. Curtain under the direction of Editor H. Kwakernaak.

† Department of Electrical Engineering, University of Minnesota, Minneapolis, MN 55455, U.S.A. Author to whom all correspondence should be addressed.

‡ Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

This paper addresses this issue of guaranteed properties for one class of gain scheduled control systems, namely, linear parameter-varying plants. This class of systems is important since it can be shown that gain scheduled control of nonlinear plants takes the form of a linear parameter-varying plant where the "parameter" is actually a reference trajectory or some endogenous signal such as the plant output (cf. Shamma and Athans, 1988, 1989). One example of a physical system whose (linearized) dynamics take the form of a parameter-varying plant is an aircraft, where the time-varying parameter is typically dynamic pressure (e.g. Stein *et al.*, 1977).

Consider a plant of the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}(\theta(t))\mathbf{x}(t) + \mathbf{B}(\theta(t))\mathbf{u}(t),$$

$$\mathbf{y}(t) = \mathbf{C}(\theta(t))\mathbf{x}(t)$$

These equations represent a linear plant whose dynamics depend on a vector of time-varying exogenous parameters, θ , which take their values in some prescribed set $\theta(t) \in \Theta$. Gain scheduled controllers for such plants typically are designed as follows. First, the designer selects a set of parameter values, $\{\theta_i\}$, which represent the range of the plant's dynamics, and designs a linear time-invariant compensator for each. Then, in between operating points, the compensators are interpolated such that for all frozen values of the parameters, the closed loop system has desirable feedback properties, such as nominal stability, robustness to unmodeled dynamics, and robust performance (Fig. 1).

Since the parameters are actually time-varying, none of these properties need carry over to the overall time-varying closed loop system. Even in the simplest case of nominal stability (i.e. no unmodeled dynamics), parameter time-variations can be destabilizing.

In this paper, conditions are given which guarantee that the closed loop system will retain the feedback properties of the frozen-time designs. These conditions formalize various heuristic ideas which have guided successful gain scheduled designs. For example, one primary guideline is "the scheduling variables should vary slowly with respect to the system dynamics." Note that this idea is simply a reminder that the original designs were based on linear time-invariant approximations to the actual plant. In this sense, these approximations must be sufficiently faithful to the true plant if one expects the global design to exhibit the desired feedback properties. In fact, it is this idea which proves most fundamental in the forthcoming analysis.

The remainder of this paper is organized as follows. This section closes with the mathematical notation to be used throughout the paper. Section 2 addresses the issues of robust stability and robust performance. The formal problem statement is given in Section 2.1. Section 2.2 presents background material on Volterra integrodifferential equations. In Section 2.3, conditions are given which guarantee time-varying robustness/performance given frozen-time robustness/performance. The conditions are presented from both a state-space and input-output viewpoint. Finally, concluding remarks are given in Section 3.

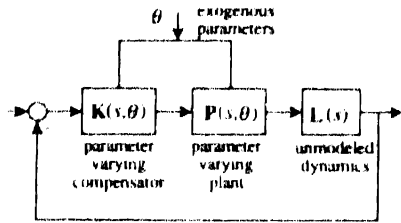


FIG. 1 A linear plant scheduling on exogenous parameters

1.2. *Mathematical notation.* Some notation regarding standard concepts for analysis of feedback systems (e.g. Desoer and Vidyasagar, 1975; Willems, 1971) is established.

\mathcal{R} denotes the field of real numbers, \mathcal{R}^+ the set $\{t \in \mathcal{R} \mid t \geq 0\}$, \mathcal{R}^n the set of $n \times 1$ vectors with elements in \mathcal{R} , and $\mathcal{R}^{n \times m}$ the set of $n \times m$ matrices with elements in \mathcal{R} . A_{ij} denotes the ij th element of the matrix A . $\|\cdot\|$ denotes both a vector norm and its induced matrix norm.

Let $f: \mathcal{R}^+ \rightarrow \mathcal{R}^n$. \hat{f} denotes the Laplace transform of f . $W_{t, \infty}$ denotes the truncation and exponential weighting operator on f defined by

$$W_{t, \infty} f(t) = \begin{cases} e^{-\alpha(t-t_0)} f(t) & t \leq T, \\ 0 & t > T. \end{cases}$$

\mathcal{L}_1 and \mathcal{L}_∞ denote the standard Lebesgue function spaces of integrable and essentially bounded measurable functions, respectively. \mathcal{M} denotes the set of measurable functions, $f: \mathcal{R}^+ \rightarrow \mathcal{R}^n$, such that

$$\|f\|_{\mathcal{M}} \stackrel{\text{def}}{=} \sup_{t \in \mathcal{R}^+} \|f(t)\| < \infty$$

$\mathcal{A}(\sigma)$ denotes the set whose elements are of the form

$$f(t) = \begin{cases} f_a(t) + \sum_{i=0}^{\infty} f_i \delta(t - t_i), & t \geq 0, \\ 0, & t < 0, \end{cases}$$

where $f_a: \mathcal{R}^+ \rightarrow \mathcal{R}^n$, $t_i \geq 0$, $f_i \in \mathcal{R}^n$, and

$$\|f\|_{\mathcal{A}(\sigma)} \stackrel{\text{def}}{=} \int_0^\infty \|f_a(t)\| e^{-\sigma t} dt + \sum_{i=0}^{\infty} \|f_i\| e^{-\sigma t_i} < \infty$$

$\mathcal{A}^{n \times m}(\sigma)$ denotes the set of $n \times m$ matrices whose elements are in $\mathcal{A}(\sigma)$. Let $\Delta \in \mathcal{A}^{n \times m}(\sigma)$ and let $\Delta' \in \mathcal{A}^{n \times m}$ be defined as $\Delta'_i = \|\Delta_i\|_{\mathcal{A}(\sigma)}$. Then $\|\Delta\|_{\mathcal{A}(\sigma)} \stackrel{\text{def}}{=} \|\Delta'\|_{\mathcal{A}(\sigma)}$ and $\mathcal{A}^{n \times m}(\sigma)$ are defined as the set of Laplace transforms of elements of $\mathcal{A}(\sigma)$ and $\mathcal{A}^{n \times m}(\sigma)$, respectively. For further details on $\mathcal{A}(\sigma)$ and $\mathcal{A}^{n \times m}(\sigma)$, see Callier and Desoer (1978) and Desoer and Vidyasagar (1975).

2. Robust stability and robust performance

2.1. *Problem statement.* Suppose that one has carried out the gain scheduled design procedure discussed in the introduction for some linear parameter-varying plant. Then for any frozen value of the parameter-vector, one has designed a feedback system which has desirable nominal stability, robust stability, and robust performance properties. Since the parameters are actually time-varying, these properties may be lost.

Now a standard practice in robust control theory is to express robust stability and robust performance requirements as the maintaining of stability in the presence of stable linear uncertainties throughout the feedback loop (e.g. Doyle, 1982; Doyle *et al.*, 1982). The original control system block diagram then may be transformed into the form of Fig. 2. In this figure, $H(\theta)$ represents a finite-dimensional parameter-varying linear system, and Δ represents a block diagonal—possibly infinite-dimensional—stable linear system which depends on only the uncertainties. In this framework, satisfying the various robust stability and robust performance specifications is equivalent to the feedback system of Fig. 2 being stable for an appropriate class of admissible

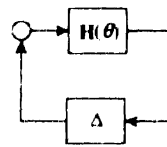


FIG. 2 General block diagram for robustness/performance analysis.

uncertainties (see Doyle, 1982; Doyle *et al.*, 1982 for details).

Employing this equivalent representation of design specifications, it then follows that the feedback diagram of Fig. 2, being a product of a gain scheduled design, is stable for all frozen parameter values. Furthermore, stability of Fig. 2 for time-varying parameters implies that the robustness and performance properties are maintained in the presence of parameter time-variations.

Let $H(\theta)$ have the following state-space realization:

$$\begin{aligned} \dot{x}(t) &= A(\theta(t))x(t) + B(\theta(t))e(t), \\ y(t) &= C(\theta(t))x(t). \end{aligned}$$

Furthermore, let the input/output relationship of Δ be given by

$$y'(t) = \int_0^t \Delta(t - \tau)y(\tau) d\tau.$$

Then the feedback equations are

$$\dot{x}(t) = A(\theta(t))x(t) + \int_0^t B(\theta(t))\Delta(t - \tau)C(\theta(\tau))x(\tau) d\tau. \tag{1}$$

This equation represents a type of linear Volterra integrodifferential equation (VIDE). In this section it is shown that the stability of (1), hence the desired robustness and performance properties, is maintained in the presence of sufficiently slow parameter time-variations. This generalizes a well known result for ordinary differential equations (e.g. Desoer, 1969).

2.2. *Volterra integrodifferential equations.* Before time-varying robustness and performance are discussed, some facts are presented regarding equations of the form in (1). Evaluating (1) along any parameter vector trajectory, one has that

$$\dot{x}(t) = A(t)x(t) + \int_0^t B(t)\Delta(t - \tau)C(\tau)x(\tau) d\tau, \tag{2}$$

where A , B and C have been appropriately redefined. This is the general form of time-varying VIDEs and will be the object of all of the forthcoming analysis. Note that any conditions imposed on (2) can be translated immediately into conditions on the parameter-varying (1).

It was stated that equation (2) falls under the class of linear VIDEs. In fact, under assumptions to be stated on Δ , (2) actually represents a combination of VIDEs and linear delay-differential equations. Thus, both types of equations are treated under the same framework. VIDEs and their stability have been studied in, for example, Burton (1983), Corduneanu and Lakshmikantham (1980), Miller (1971), and delay-differential equations in Corduneanu and Lakshmikantham (1980), Driver (1977) and Hale (1977).

In this section, assumptions on (2) are given, a definition of exponential stability is introduced, a sufficient condition for exponential stability in the case of time-invariant A , B and C matrices is given, and a perturbational result time-invariant VIDEs is presented.

Consider the VIDE

$$\dot{x}(t) = A(t)x(t) + \int_0^t B(t)\Delta(t - \tau)C(\tau)x(\tau) d\tau, \quad t > t_0, \tag{3}$$

with initial condition

$$\begin{cases} x(t) = \phi(t), & 0 \leq t \leq t_0, \\ x(t_0^+) = \phi(t_0). \end{cases} \quad \phi \in \mathcal{R}; \tag{4}$$

Note that an initial condition for (3) consists of both an initial time, t_0 , and an initial function, ϕ .

The following assumptions are made on (3):

Assumption 1. The matrices $A: \mathcal{R}^n \rightarrow \mathcal{R}^{n \times n}$, $B: \mathcal{R}^n \rightarrow \mathcal{R}^{n \times m}$ and $C: \mathcal{R}^n \rightarrow \mathcal{R}^{p \times n}$ are globally bounded and Lipschitz continuous with Lipschitz constants L_A , L_B , and L_C , respectively.

Assumption 2. For some $\sigma \geq 0$, $\Delta \in \mathcal{M}^{m \times p}(-\sigma)$.

VIDEs containing an integral operator as in assumption 2 have been studied in Corduneanu (1973), Corduneanu and Luca (1975) and Luca (1979) and references are contained in Corduneanu and Lakshmikantham (1980). In case of time-invariant A , B and C matrices, solutions to (3) can be explicitly characterized as follows:

Theorem 1. Consider the time-invariant VIDE

$$\dot{x}(t) = Ax(t) + \int_0^t B\Delta(t-\tau)Cx(\tau) d\tau + f(t), \quad t > t_0, \quad f \in \mathcal{L}_\infty, \quad (5)$$

with initial condition (4) under assumption 2. The unique solution to (5) is given by

$$x(t+t_0) = R(t)x(t_0) + \int_0^t R(t-\tau)(f(\tau+t_0) + F(\tau+t_0)) d\tau, \quad t > 0,$$

where

$$F(t+t_0) \triangleq \int_0^{t_0} B\Delta(t+t_0-\tau)C\phi(\tau) d\tau, \quad t > 0,$$

and R , known as the resolvent matrix, is the unique matrix satisfying

$$R(t) = I + \int_0^t (AR(\tau) + \int_0^\tau B\Delta(\tau-\xi)CR(\xi) d\xi) d\tau, \quad t > 0, \quad R(0^+) = I.$$

Proof. See Corduneanu (1973) and Corduneanu and Luca (1975). ■

A definition of exponential stability for VIDEs is now introduced.

Definition 1. Consider the VIDE (3) with initial condition (4). This VIDE is said to be exponentially stable if there exists constants m , λ , $\beta > 0$ where $\beta \geq \lambda$ such that

$$\|x(t)\| \leq me^{-\lambda(t-t_0)} \|W_{t_0, \beta} \phi\|_\infty, \quad \forall t \geq t_0.$$

It is stressed that the constants m , λ and β are independent of the initial condition (ϕ , t_0).

This definition implies that not only does the state decay exponentially, but also with a magnitude which is proportional to an exponentially forgotten initial function. The convention that $\beta \geq \lambda$ implies that the solutions cannot decay faster than they are forgotten. Furthermore, this inequality will be needed in subsequent proofs (cf. Theorem 2).

The following theorem gives a sufficient condition for exponential stability for time-invariant VIDEs.

Theorem 2. Consider the time-invariant VIDE (5) with initial condition (4). A sufficient condition for exponential stability is that there exist a constant $\beta > 0$ such that

$$s \mapsto (sI - A - B\hat{\Delta}(s)C)^{-1} \in \mathcal{M}^{n \times n}(-2\beta), \quad (6)$$

$$\hat{\Delta} \in \mathcal{M}^{m \times p}(-2\beta). \quad (7)$$

Proof. It is first shown that the resolvent matrix R is bounded by a decaying exponential. From the definition of R in Theorem 1, one has that R satisfies almost everywhere

$$\dot{R}(t) = AR(t) + \int_0^t B\Delta(t-\tau)CR(\tau) d\tau, \quad t > 0. \quad (8)$$

Taking the Laplace transform of (8) shows that

$$\hat{R}(s) = (sI - A - B\hat{\Delta}(s)C)^{-1}.$$

It follows by hypothesis that $R \in \mathcal{M}^{n \times n}(-2\beta)$. Furthermore, it may be seen from the definition of R in Theorem 1 that R contains no impulses, hence $R \in \mathcal{L}_1$. From (8), it follows that $R \in \mathcal{L}_1$. These two imply that $R \in \mathcal{L}_\infty$. Now define

$$R'(t) \triangleq R(t)e^{\beta t}.$$

Then $R' \in \mathcal{M}^{n \times n}(-\beta)$ because $R \in \mathcal{M}^{n \times n}(-2\beta)$. Using the same arguments as above along with

$$\dot{R}(t) = (A + \beta I)R'(t) + \int_0^t B\Delta(t-\tau)e^{\beta(t-\tau)}CR'(\tau) d\tau, \quad t > 0,$$

it follows that R' and $\dot{R}' \in \mathcal{L}_1$, hence $R' \in \mathcal{L}_\infty$. Thus, it follows that there exists a constant k_1 , for example $k_1 = \|R'\|_{\mathcal{L}_\infty}$, such that

$$\|R(t)\| \leq k_1 e^{-\beta t}, \quad t > 0.$$

Now recall that the solution to (5) is given by

$$x(t+t_0) = R(t)x(t_0) + \int_0^t R(t-\tau)F(\tau+t_0) d\tau, \quad t > 0,$$

where

$$F(t+t_0) \triangleq \int_0^{t_0} B\Delta(t+t_0-\tau)C\phi(\tau) d\tau, \quad t > 0.$$

It is now shown that F is also bounded by a decaying exponential. Rewriting the definition of F ,

$$\begin{aligned} F(t+t_0) &= \int_0^{t_0} B\Delta(t+t_0-\tau)e^{\beta(t+t_0-\tau)}C e^{-\beta(t+t_0-\tau)}\phi(\tau) d\tau \\ &\leq e^{-\beta t} \int_0^{t_0} B\Delta(t+t_0-\tau)e^{\beta(t_0-\tau)}C e^{-\beta(t_0-\tau)}\phi(\tau) d\tau. \end{aligned}$$

Since $\Delta \in \mathcal{M}^{m \times p}(-2\beta)$, it follows that there exists a constant k_2 , for example $k_2 = \|B\| \|\Delta\|_{\mathcal{M}^{m \times p}} \|C\|$, such that

$$\|F(t+t_0)\| \leq k_2 e^{-\beta t} \|W_{t_0, \beta} \phi\|_\infty.$$

Using the exponential bounds on R and F to bound x ,

$$\begin{aligned} \|x(t+t_0)\| &\leq k_1 e^{-\beta t} \|x(t_0)\| + \int_0^t k_1 e^{-\beta(t-\tau)} k_2 e^{-\beta \tau} \|W_{t_0, \beta} \phi\|_\infty d\tau \\ &\leq k_1 e^{\beta t} (1 + k_2 t) \|W_{t_0, \beta} \phi\|_\infty \\ &\leq k_1 \left(1 + \frac{2k_2}{\beta} e^{\beta t}\right) e^{-\beta t/2} \|W_{t_0, \beta} \phi\|_\infty, \end{aligned}$$

which completes the proof. ■

Theorem 2 is novel in that it takes a state-space approach, rather than an input/output approach, to the robust stability of time-invariant linear systems. This approach is chosen since it corresponds to the original motivation of parameter-varying gain scheduled systems. Nevertheless, standard results on robust stability can be obtained from this theorem. Rewriting $\hat{R}(s)$, one has that

$$\hat{R}(s) = (I - (sI - A)^{-1} B\hat{\Delta}(s)C)^{-1} (sI - A)^{-1}.$$

Now suppose that A is a stable matrix; thus $s \mapsto (sI - A)^{-1} \in \mathcal{M}^{n \times n}(-2\beta)$ for some $\beta > 0$. Assume further that $\hat{\Delta} \in \mathcal{M}^{m \times p}(-2\beta)$. Then

$$s \mapsto (I - (sI - A)^{-1} B\hat{\Delta}(s)C) \in \mathcal{M}^{n \times n}(-2\beta).$$

Under these conditions, $\hat{R} \in \mathcal{M}^{n \times n}(-2\beta)$ if (Dewoer and Vidyasagar, 1975)

$$\begin{aligned} \inf_{\Re s > -2\beta} |\det(I - (sI - A)^{-1} B\hat{\Delta}(s)C)| \\ = \inf_{\Re s > -2\beta} |\det(I - C(sI - A)^{-1} B\hat{\Delta}(s))| > 0. \end{aligned}$$

However, a sufficient condition for the above equation is that

$$\|C(-2\beta + j\omega)I - A)^{-1} B\hat{\Delta}(-2\beta + j\omega)\| \leq \gamma < 1, \quad \forall \omega \in \mathcal{R}.$$

As $\beta \rightarrow 0$, this condition approaches the standard small-gain robustness condition for time-invariant linear systems (e.g. Chen and Dewoer, 1982; Doyle and Stein, 1981). Unlike

previous results, however, Theorem 2 gives a quantitative indication of the rate of exponential decay of the state-variables.

This section concludes with a presentation of a perturbational result for time-invariant VIDEs.

Theorem 3. Consider the following perturbation of the VIDE (5):

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \int_0^t \mathbf{B}\Delta(t-\tau)\mathbf{C}\mathbf{x}(\tau) d\tau + (\mathbf{g}\mathbf{x})(t), \quad t \geq t_0, \quad (9)$$

where \mathbf{g} is an integral operator on \mathbf{x} . Let

$$\begin{aligned} s \mapsto (s\mathbf{I} - \mathbf{A} - \mathbf{B}\hat{\Delta}(s)\mathbf{C})^{-1} &\in \mathcal{A}^{m \times n}(-2\beta), \\ \hat{\Delta} &\in \mathcal{A}^{m \times p}(-2\beta), \end{aligned}$$

for some $\beta > 0$. Assume further that there exist constants $k > 0$ and $\alpha \neq \beta$ such that

$$\|(\mathbf{g}\mathbf{x})(t)\| \leq k \|\mathcal{W}_{t_0, \beta}\mathbf{x}\|_{\mathcal{B}}, \quad \forall t \geq 0, \quad \forall \mathbf{x} \in \mathcal{B}.$$

Under these conditions, there exists a $\gamma > 0$ such that the VIDE (9) is exponentially stable for $k < \gamma$.

Proof. Let (9) have an initial condition (4). Define $\mathbf{z}(t) \stackrel{\text{def}}{=} \mathbf{x}(t + t_0)$. As in Theorem 1, one has that

$$\begin{aligned} \mathbf{z}(t) = \mathbf{R}(t)\mathbf{z}(t_0) + \int_0^t \mathbf{R}(t-\tau)(\mathbf{F}(\tau + t_0) \\ + (\mathbf{g}\mathbf{x})(\tau + t_0)) d\tau, \quad t \geq 0, \end{aligned}$$

where \mathbf{R} is the resolvent matrix and

$$\mathbf{F}(t + t_0) = \int_0^{t_0} \mathbf{B}\Delta(t + t_0 - \tau)\mathbf{C}\phi(\tau) d\tau, \quad t \geq 0.$$

As in the proof of Theorem 2, there exist k_1 and k_2 such that

$$\begin{aligned} \|\mathbf{R}(t)\| &\leq k_1 e^{-\beta t}, \\ \|\mathbf{F}(t + t_0)\| &\leq k_2 e^{-\beta t} \|\mathcal{W}_{t_0, \beta}\phi\|_{\mathcal{B}}. \end{aligned}$$

Using these bounds to bound \mathbf{z} ,

$$\begin{aligned} \|\mathbf{z}(t)\| &\leq k_1 e^{-\beta t} \|\mathbf{z}(0)\| + \int_0^t k_1 e^{-\beta(t-\tau)} \\ &\quad \times (k_2 e^{-\beta \tau} \|\mathcal{W}_{t_0, \beta}\phi\|_{\mathcal{B}} + k \|\mathcal{W}_{\tau+t_0, \beta}\mathbf{x}\|_{\mathcal{B}}) d\tau. \end{aligned}$$

By definition of the \mathcal{W} operator

$$\begin{aligned} \|\mathcal{W}_{\tau+t_0, \beta}\mathbf{x}\|_{\mathcal{B}} &\leq \|\mathcal{W}_{\tau, \beta}\mathbf{x}\|_{\mathcal{B}} \stackrel{\text{def}}{=} \sup_{\xi \in [0, \tau+t_0]} |e^{-\beta(\tau+t_0-\xi)} \mathbf{x}(\xi)| \\ &\leq e^{-\beta \tau} \left(\sup_{\xi \in [0, t_0]} |e^{-\beta(t_0-\xi)} \phi(\xi)| + \sup_{\xi \in [t_0, \tau+t_0]} |e^{-\beta(t_0-\xi)} \mathbf{x}(\xi)| \right). \end{aligned}$$

Thus,

$$\begin{aligned} e^{\beta t} \|\mathbf{z}(t)\| &\leq k_1 (1 + (k + k_2)t) \|\mathcal{W}_{t_0, \beta}\phi\|_{\mathcal{B}} \\ &\quad + \int_0^t k_1 k \sup_{\xi \in [0, \tau]} |e^{\beta \xi} \mathbf{z}(\xi)| d\tau. \end{aligned}$$

Since the right-hand side of the above equation is a nondecreasing function of time,

$$\begin{aligned} \sup_{\xi \in [0, t]} |e^{\beta \xi} \mathbf{z}(\xi)| &\leq k_1 (1 + (k + k_2)t) \|\mathcal{W}_{t_0, \beta}\phi\|_{\mathcal{B}} \\ &\quad + \int_0^t k_1 k \sup_{\xi \in [0, \tau]} |e^{\beta \xi} \mathbf{z}(\xi)| d\tau. \end{aligned}$$

Rewriting this equation yields

$$f(t) \leq \kappa_1 + \kappa_2 t + \kappa_3 \int_0^t f(\tau) d\tau,$$

where κ_1 , κ_2 and f are defined in the obvious manner. Applying the Bellman–Gronwall inequality (e.g. Desoer and Vidyasagar, 1975),

$$f(t) \leq \left(\kappa_1 + \frac{\kappa_2}{\kappa_1} \right) e^{\kappa_2 t} - \frac{\kappa_2}{\kappa_1}.$$

Thus,

$$\begin{aligned} \|\mathbf{z}(t)\| &\leq \left(\left(k_1 + 1 + \frac{k_2}{k} \right) e^{-(\beta - k_1 k)t} - \left(1 + \frac{k_2}{k} \right) e^{-\beta t} \right) \|\mathcal{W}_{t_0, \beta}\phi\|_{\mathcal{B}} \\ &= \left(k_1 + \left(\frac{k + k_2}{k} \right) (1 - e^{-(\beta - k_1 k)t}) \right) e^{-(\beta - k_1 k)t} \|\mathcal{W}_{t_0, \beta}\phi\|_{\mathcal{B}} \\ &\leq k_1 (1 + (k_2 + k)t) e^{-(\beta - k_1 k)t} \|\mathcal{W}_{t_0, \beta}\phi\|_{\mathcal{B}}. \end{aligned}$$

Thus, it is seen that $k < \gamma \stackrel{\text{def}}{=} \beta/k_1$ implies exponential stability. Furthermore,

$$\|\mathbf{z}(t)\| \leq k_1 \left(1 + \frac{2(k_2 + k)}{e(\beta - k_1 k)} \right) e^{-(\beta - k_1 k)t/2} \|\mathcal{W}_{t_0, \beta}\phi\|_{\mathcal{B}}. \quad (10)$$

2.3. Robustness and performance of slowly-varying linear systems. In this section, it is shown that if the time-varying VIDE (3) is exponentially stable for all frozen values of time, then it is exponentially stable for sufficiently slow time-variations. In terms of the original motivation of guaranteed properties of gain scheduled control systems, this means that robust stability and robust performance are maintained provided that parameter variations are sufficiently slow.

Before proceeding with the main theorem, some assumptions and definitions are given.

Assumption 3. Consider the time-varying VIDE (3) under assumptions 1–2. The matrix functions \mathbf{A} , \mathbf{B} , \mathbf{C} and Δ are such that for each $\xi \in \mathcal{H}'$,

$$\begin{aligned} s \mapsto (s\mathbf{I} - \mathbf{A}(\xi) - \mathbf{B}(\xi)\hat{\Delta}(s)\mathbf{C}(\xi))^{-1} &\in \mathcal{A}^{n \times n}(-2\beta), \\ \hat{\Delta} &\in \mathcal{A}^{m \times p}(-2\beta). \end{aligned}$$

Via Theorem 2, these two conditions imply that for each $\xi \in \mathcal{H}'$, the time-invariant VIDE

$$\dot{\mathbf{x}}(t) = \mathbf{A}(\xi)\mathbf{x}(t) + \int_0^t \mathbf{B}(\xi)\Delta(t-\tau)\mathbf{C}(\xi)\mathbf{x}(\tau) d\tau, \quad t \geq t_0, \quad (11)$$

is exponentially stable as in Definition 1. It is further assumed that this exponential stability is uniform in ξ . That is, there exist exponential stability constants (m, λ, β) for (11) which are independent of ξ .

Note that in the case of gain scheduling, one may use Theorem 2 to verify exponential stability for all frozen parameter values, hence for all time. Furthermore, if the parameters take their values in some compact set, then the exponential stability is uniform.

Definition 2. Let assumptions 1–2 hold. Under assumption 1, let k_B and k_C satisfy $\forall \xi \in \mathcal{H}'$

$$\|\mathbf{B}(\xi)\| \leq k_B, \quad \|\mathbf{C}(\xi)\| \leq k_C.$$

Then the measure, K , of the rate of time-variations of (3) is defined as

$$K \stackrel{\text{def}}{=} L_A + L_B \|\Delta\|_{\mathcal{A}(t, \beta)} k_C + k_B \|\Delta\|_{\mathcal{A}(t, \beta)} L_C.$$

The question of slowly time-varying stability of linear VIDE's is now addressed.

Theorem 4. Consider the time-varying VIDE (3) under assumptions 1–3. Under these conditions, (3) is exponentially stable for sufficiently small K , or equivalently, for sufficiently slow time-variations in \mathbf{A} , \mathbf{B} and \mathbf{C} .

Proof. Let \mathbf{R}_ξ denote the resolvent matrix associated with the frozen-time VIDE (11). From assumption 2 of uniform frozen-time exponential stability, there exists a constant K_1 such that

$$\|\mathbf{R}_\xi(t)\| \leq K_1 e^{-\beta t}, \quad \forall \xi \in \mathcal{H}'.$$

Similarly, define

$$K_2 \stackrel{\text{def}}{=} k_B \|\Delta\|_{\mathcal{A}(t, \beta)} k_C.$$

Note that the constants K_1 and K_2 represent worst case values of their frozen- ξ analogs in the proof of Theorem 2.

Let the initial condition of (3) be (4), and let t_n denote $t_n + nT$, where T is some constant interval to be chosen. Approximating A , B and C by piecewise constant matrices, one has that

$$\begin{aligned} \dot{x}(t) = & A(t_n)x(t) + \int_{t_n}^t B(t_n)\Delta(t-\tau)C(t_n)x(\tau) d\tau \\ & + \int_{t_n}^t B(t)\Delta(t-\tau)C(\tau)x(\tau) d\tau + (g_n x)(t), \end{aligned}$$

where

$$\begin{aligned} (g_n x)(t) = & (A(t) - A(t_n))x(t) \\ & + \int_{t_n}^t (B(t) - B(t_n))\Delta(t-\tau)C(t_n)x(\tau) d\tau \\ & + \int_{t_n}^t B(t)\Delta(t-\tau)(C(\tau) - C(t_n))x(\tau) d\tau. \end{aligned}$$

Then

$$\|(g_n x)(t)\| \leq KT \|W_{t_n, \beta} x\|_{\infty}, \quad t_n \leq t \leq t_{n+1}$$

Choose any $\eta \in (0, \beta)$. Then using arguments exactly parallel to those of the proof of Theorem 3, it can be shown that

$$KT \leq \frac{\beta - \eta}{2K_1} \quad (12)$$

implies that

$$\begin{aligned} \|x(t)\|_{\infty} \leq & K_1 \left(1 + \frac{2(K_1 + \frac{\beta - \eta}{2K_1})}{c(\beta - K_1 - \frac{\beta - \eta}{2K_1})} \right) e^{-(\beta - K_1)(\beta - \eta)(2K_1 + 2)(t - t_n)} \|W_{t_n, \beta} x\|_{\infty} \\ & + M e^{-(\beta - \eta)(4)(t - t_n)} \|W_{t_n, \beta} x\|_{\infty}, \end{aligned}$$

where M is defined in the obvious manner [cf. (10)]. Note that since $K_1 \geq 1$ by definition, one has $M \geq 1$ in general. This bound on x further implies that

$$\begin{aligned} \|W_{t_n, \beta} x\|_{\infty} & \stackrel{\text{def}}{=} \sup_{\xi \in [0, t_n]} \|e^{-(\beta - K_1)(t_n - \xi)} x(\xi)\| \\ & = \max \left(\sup_{\xi \in [0, t_n]} \|e^{-(\beta - K_1)(t_n - \xi)} x(\xi)\|, \sup_{\xi \in (t_n, t_{n+1})} \|e^{-(\beta - K_1)(t_n - \xi)} x(\xi)\| \right) \\ & \leq \max \left(e^{-\beta T} \sup_{\xi \in [0, t_n]} \|e^{-(\beta - K_1)(t_n - \xi)} x(\xi)\|, \right. \\ & \quad \left. \sup_{\xi \in (t_n, t_{n+1})} e^{-(\beta - K_1)(t_n - \xi)} M e^{-(\beta - \eta)(4)(t_n - \xi)} \|W_{t_n, \beta} x\|_{\infty} \right) \\ & = \max \left(e^{-\beta T} \|W_{t_n, \beta} x\|_{\infty}, \right. \\ & \quad \left. \sup_{\xi \in (t_n, t_{n+1})} M e^{(\beta - (\beta - \eta)(4))t_n} e^{-(\beta - K_1)(t_n - \xi)} e^{-(\beta - \eta)(4)(t_n - \xi)} \|W_{t_n, \beta} x\|_{\infty} \right) \\ & = M e^{-(1/2)(\beta + \eta)(2)T} \|W_{t_n, \beta} x\|_{\infty} \end{aligned}$$

In order to guarantee (12), choose

$$T = 4 \ln M / (\beta - \eta)$$

Then

$$K \leq \frac{(\beta - \eta)^2}{8K_1 \ln M}$$

implies the desired (12). Furthermore, recursively applying the bound on $\|W_{t_n, \beta} x\|_{\infty}$ shows that

$$\begin{aligned} \|x(t)\| & \leq M e^{-(\beta + \eta)(4)(t - t_n)} \|W_{t_n, \beta} x\|_{\infty} \\ & \leq M e^{-(1/2)(\beta + \eta)(2)(t - t_0 - nT)} \\ & \quad \times (M e^{-(1/2)(\beta + \eta)(2)T})^n \|W_{t_0, \beta} x\|_{\infty} \\ & = M e^{-(\eta/2)(t - t_0)} e^{-(1/2)(\beta - \eta)(2)(t - t_0)} \\ & \quad \times (M e^{-(1/2)(\beta - \eta)(2)T})^n \|W_{t_0, \beta} x\|_{\infty} \end{aligned}$$

Substituting the choice of T into the above equation yields

$$\|x(t)\| \leq M e^{-(\eta/2)(t - t_0)} \|W_{t_0, \beta} x\|_{\infty},$$

which completes the proof. ■

The main idea behind Theorem 4 is as follows. The time-varying VIDE (3) is approximated by a piecewise constant VIDE which is piecewise exponentially stable. Thus on each interval, the time-varying VIDE is decomposed into a constant part and a time-varying perturbation. Using Theorem 3, the solution will decay provided that the piecewise constant approximations are sufficiently accurate over sufficiently long intervals, which, in turn, is guaranteed by sufficiently slow time-variations.

3. Concluding remarks

This paper has addressed the robust stability and robust performance of parameter-varying linear systems in the context of gain-scheduling. The results may be summarized as follows. Essentially, it was shown that a gain scheduled system which has desirable feedback properties for all frozen values of the parameters maintains these properties provided that the parameter time-variations are sufficiently slow. Explicit sufficient conditions on the parameter time-variations were given. Thus the heuristic guidelines of "scheduling on a slow variable" has been transformed into quantitative statements.

Unfortunately, the actual bounds on the parameter time-variations may be difficult—at best—to compute. For example, to verify the sufficient conditions for frozen-parameter exponential stability (cf. Theorem 2) would require satisfying a small-gain condition off the $j\omega$ -axis. Once these conditions are verified, one can then use Theorem 4 to guarantee time-varying stability. However, this requires computation of the measure of time-variations in Definition 2 (K), a bound on the resolvent matrix for the frozen-parameter systems (K_1), and a bound on the exponentially weighted input/output norm of the linear uncertainties ($\|\Delta\|_{\infty, \beta, \eta}$). Furthermore, even if verified these results are apt to be conservative.

In spite of these limitations, the value of the results is that they lead to new insights into gain-scheduled systems. For example, in performing the frozen parameter designs, one must guarantee some degree of internal exponential stability in addition to the input/output stability of standard robustness tests. Furthermore, the sufficiency of the conditions is simply a reminder that the designs were based on time-invariant approximations to the actual time-varying plant. If these approximations are inaccurate, then one should not demand guarantees on the overall gain scheduled system.

Acknowledgement—This research was supported by the NASA Ames and Langley Research Centers under grant NASA/NAG 2-297.

References

- Burton, T. A. (1983). *Volterra Integral and Differential Equations*. Academic Press, New York.
- Callier, F. M. and C. A. Desoer (1978). An algebra of transfer functions of distributed linear time-invariant systems. *IEEE Trans. Circ. Syst.*, **CAS-25**, 651–662.
- Chen, M. J. and C. A. Desoer (1982). Necessary and sufficient condition for robust stability of linear distributed feedback systems. *Int. J. Control*, **35**, 255–267.
- Corduneanu, C. (1973). Some differential equations with delay. *Proc. Equadiff 3 (Czechoslovak Conference on Differential Equations and their Applications)*, Brno, pp. 105–114.
- Corduneanu, C. and V. Lakshmikantham (1980). Equations with unbounded delay: A survey. *Nonlinear Anal. Theory Methods Appl.*, **4**, 831–877.
- Corduneanu, C. and N. Luca (1975). The stability of some feedback systems with delay. *J. Math. Anal. Appl.*, **51**, 377–393.
- Desoer, C. A. (1969). Slowly varying system $\dot{x} = A(t)x$. *IEEE Trans. Aut. Control*, **AC-14**, 780–781.
- Desoer, C. A. and M. Vidyasagar (1975). *Feedback Systems: Input-Output Properties*. Academic Press, New York.
- Doyle, J. (1982). Analysis of feedback systems with structural uncertainties. *Proc. IEE* **129**, Part D, 242–250.

- Doyle, J. C. and G. Stein (1981). Multivariable feedback design: concepts for a classical/modern synthesis. *IEEE Trans. Aut. Control*, **AC-26**, 4-16.
- Doyle, J. C., J. E. Wall and G. Stein (1982). Performance and robustness analysis for structural uncertainty. *Proc. 21st Conf. Decision and Control*, pp 629-636.
- Driver, R. D. (1977) *Ordinary and Delay Differential Equations*. Springer, New York.
- Hale, J. (1977). *Theory of Functional Differential Equations*. Springer, New York.
- Luca, N. (1979). The stability of the solutions of a class of integrodifferential systems with infinite delays. *J. Math. Anal. Applics.*, **67**, 323-339.
- Miller, R. K. (1971). *Nonlinear Volterra Integral Equations*. Benjamin, Menlo Park, CA.
- Shamma, J. S. and M. Athans (1988). Guaranteed properties of nonlinear gain scheduled control systems. *Proc. 27th IEEE Conf. on Decision and Control*.
- Shamma, J. S. and M. Athans (1990). Analysis of nonlinear gain scheduled control systems, *IEEE Trans. Aut. Control*, **35**, 898-907.
- Stein, G., G. L. Hartmann and R. C. Hendrick (1977). Adaptive control laws for F-8 Flight Test. *IEEE Trans. Aut. Control*, **AC-22**, 758-767.
- Stein, G. (1980). Adaptive flight control—a pragmatic view. In Narendra, K. S. and R. V. Monopoli (Eds), *Applications of Adaptive Control*. Academic Press, New York.
- Willems, J. C. (1971). *The Analysis of Feedback Systems*. MIT Press, Cambridge, MA.

Brief Paper

Arbitrarily Small Sensitivity in Multiple-input-output Uncertain Feedback Systems*

ODED YANIV†

Key Words—Control theory; feedback control; multivariable control systems; robust control; frequency domain.

Abstract—A square, linear time invariant, multiple-input multiple-output plant R is considered, known only to belong to a set $\{R\}$. The plant is embedded in a feedback structure designed such that the closed loop response belongs to a specified set for all $R \in \{R\}$. This paper develops sufficient conditions on the uncertain set $\{R\}$ such that arbitrarily small sensitivity and internal stability can be achieved by the Horowitz synthesis method. Specifically, we show that, in addition to generalized single-input single-output like conditions, there exists a condition on the pole/zero excess of the controller.

Abbreviations—SISO, single-input single-output; MIMO, multi-input multi-output; LTI, linear time invariant; TF, transfer function; MTF, matrix transfer function; RHP, closed right half plane; RRHP, regular in the RHP; HFG, high frequency gain.

1. Introduction

THE TWO main reasons for using feedback are to reduce the closed loop sensitivity to plant uncertainty and to meet disturbance rejection specifications. This raises the question of whether or not there exists a common LTI stabilization controller for a given family of plants which also achieves *a priori* closed loop performance. The stabilization problem appears in the literature as the *simultaneous stabilization problem* or the *robust design problem*, and has been discussed by many authors (for example Vidyasagar, 1985; Saberi, 1985; Ghosh, 1986; Schmitendorf and Holot, 1988; Barmish, 1989; Wei and Barmish, 1988) who include a very good introduction to the problem. A more difficult problem is the simultaneous stabilization problem with constraints on the closed loop system's performance. This problem appears in many applications, for example in flight control (Garnell and East, 1977), large space structures (Wei and Byun, 1989), robotics (Paul, 1981) and chemical plant processing (Skogestad *et al.*, 1988). This paper is devoted to the problem of robust stabilization with constraints. Explicit criteria are established on the set $\{R\}$ under which a controller exists for arbitrary robustness and performance. Specifically, the elements of R^{-1} must be stable, the sign of the high frequency gain of the leading principal $(k-1)$ minor of R divided by the leading principal minor k is fixed over $\{R\}$ for $k=1, \dots, n$, and there exists a condition on the pole/zero excess of the controller. The Horowitz design method for MIMO systems is the framework for this work (Yaniv and Horowitz, 1986).

2. Statement of the problem

In Fig. 1 $R = [R_{ij}(s)]$ is the plant model, an $n \times n$ matrix

* Received 4 April 1989; revised 19 February 1990; revised 6 July 1990; received in final form 4 August 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor P. Guimaraes Feireira under the direction of Editor H. Kwakernaak.

† Faculty of Engineering, Department of Electrical Systems, Tel-Aviv University, Tel-Aviv 69978, Israel.

whose elements are proper finite rational TF. Due to uncertainty in the plant's parameters, $R \in \{R\}$ where $\{R\}$ is a finite set of possible LTI plants. For all $i, j = 1, 2, \dots, n$ let the following definitions hold:

- $A_{ij}(\omega) \approx$ a non-negative function of ω
- $B_{ij}(\omega) \approx$ a positive function of ω , $B_{ij}(\omega) \geq A_{ij}(\omega)$
- $V_{ij}(\omega) \approx$ a positive function of ω
- t_{ij} = the system's TF from input r_j to output y_i
- d_{ij} = the system's TF from input r_i to output y_i .

Problem 1 Find strictly proper elements of the MTF $F = [f_{ij}(s)]$ and $G = \text{diag}[g_i(s)]$ in Fig. 1, so that the following condition are satisfied for all $R \in \{R\}$:

Stability: The closed system shall be internally stable. (2.1)

Closed loop performance: For a given ω_0 , A_{ij} , B_{ij} and V_{ij}

$$A_{ij}(\omega) \approx |t_{ij}(i\omega)| \approx B_{ij}(\omega), |d_{ij}(i\omega)| \approx V_{ij}(\omega); \quad \text{for } \omega \approx \omega_0 \text{ and } i, j = 1, \dots, n \quad (2.2)$$

The frequency ω_0 is defined as the frequency above which the disturbances are low in magnitude and sensitivity to plant variation is not important (because at sufficiently high frequencies the benefits of feedback are negligible). In SISO systems ω_0 is often chosen as the frequency, for which the open loop is about -3db .

This problem has been solved by Yaniv and Horowitz (1986) who developed a synthesis procedure.

Definition 1. An *arbitrarily small sensitivity set* is a set $\{R\}$ such that for any choice of $A_{ij}(\omega)$, $B_{ij}(\omega)$, $V_{ij}(\omega)$ and frequency ω_0 there exists a solution F, G to Problem 1.

Problem 2. Under the constraints of Problem 1, find classes of arbitrarily small sensitivity sets to which the Horowitz synthesis method for MIMO systems is applicable.

This paper is devoted to Problem 2.

Assumption 1. The design parameter G is diagonal, G^{-1} is RRHP, and for all $R \in \{R\}$, R is proper, R^{-1} exists and is RRHP. If G is minimum phase, as is the case in most applications, then clearly G^{-1} is RRHP. A similar statement cannot be made about R and R^{-1} .

3. Internal stability

3.1. Brief review of the Horowitz method The feedback system of Fig. 1 is described by:

$$[I + G^{-1}R^{-1}][t_{ij}] \approx [f_{ij}]; \quad T \approx [t_{ij}]. \quad (3.1)$$

This MIMO problem can be reduced to n^2 SISO problems using the Gauss's elimination method (Gantmacher, 1960). Applying this elimination method to equation (3.1) gives for $k=1, \dots, n$:

$$\sum_{i=k+1}^n \left(\delta_{ik} + \frac{P_{ik}^*}{g_i} \right) t_{ij} \approx f_{ij} - \sum_{i=1}^{k-1} \frac{P_{ik}^* f_{ij}}{g_i}; \quad i \geq k \quad (3.2)$$

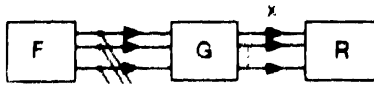


FIG. 1. n -input n -output feedback structure. R is an uncertain plant.

where $\delta_{iu} = 1$ if $i = u$ and $\delta_{iu} = 0$ if $i \neq u$, and P_{ij}^k satisfy the recursive equation:

$$[P_{ij}^k] = R^{-1}; \quad P_{ij}^{k+1} = P_{ij}^k - \frac{P_{ik}^k P_{kj}^k}{P_{kk}^k + g_k}; \quad i, j > k. \quad (3.3)$$

Substituting $i = k$ in equation (3.2) gives the following n equations:

$$\sum_{u=k+1}^n \left(\delta_{ku} + \frac{P_{ku}^k}{g_k} \right) t_{u1} = f_{k1} \\ \sum_{u=k+1}^n \frac{P_{ku}^k}{g_k} f_{u1} = f_{k1}^0; \quad k = 1, \dots, n. \quad (3.4)$$

Equation (3.4) is a matrix equation of the form:

$$[h_{ij}](t_{11}, \dots, t_{n1})' = (f_{11}^0, \dots, f_{n1}^0) \quad (3.5)$$

Rearranging the terms in (3.4) gives:

$$\frac{L_k}{1+L_k} f_{k1} = \sum_{j=1}^k \frac{P_{kj}^k q_k}{1+L_k} f_{uj} = \sum_{j=1}^k \frac{P_{kj}^k q_k}{1+L_k} \quad (3.6)$$

where

$$L_k = g_k / P_{kk}^k; \quad q_k = 1/P_{kk}^k \quad (3.7)$$

Therefore by equation (3.6), the MIMO problem is reduced to n^2 SISO problems as follows. In step k the designer finds g_k and f_{k1} , which are the only unknown parameters in inequality (3.8) below, so that $(1+L_k)$ is stable and:

$$A_{k1} = \frac{f_{k1} L_k}{1+L_k} - \frac{D_{k1} q_k}{1+L_k} \leq B_{k1} \quad (3.8)$$

where

$$D_{k1} = \sum_{u=1, \dots, k-1} P_{ku}^k f_{u1} - \sum_{u=k+1, \dots, n} P_{ku}^k t_{u1} \quad (3.9)$$

and the t_{u1} appearing in D_{k1} can have all values that will satisfy

$$A_{u1} \leq |t_{u1}| \leq B_{u1}. \quad (3.10)$$

The problem defined above to be solved in step k is a SISO problem. It has a solution if certain conditions are satisfied (Horowitz, 1979, Appendix 1). At the end of the design process, the synthesized $G = \text{diag}[g_i]$ and $F = [f_{ij}]$ guarantee that (2.2) is met. A modification of the Horowitz method for achieving the disturbance specifications V_y in (2.2) is given in Yaniv and Horowitz (1986). Based on the above synthesis procedure, conditions are now developed under which it generates an internally stable solution.

3.2. Conditions for internal stability.

Lemma 1. Under Assumption 1, if for all $k = 1, \dots, n$ $(g_k + P_{kk}^k)$ has no RHP zeros, then P_{kk}^k are RRHP and the elements of $T = [t_{ij}]$ are RRHP.

Proof. By induction on equation (3.3), P_{kk}^k are RRHP, which guarantees that h_{ij} and f_{ij}^0 in equation (3.5) are RRHP. Since $[h_{ij}]$ is upper triangular $[h_{ij}]^{-1}$ is also RRHP. Hence the elements of $[t_{ij}] = [h_{ij}]^{-1}[f_{ij}^0]$ are RRHP.

Theorem 1. Suppose that the conditions of Lemma 1 are satisfied; then the system described by equation (3.1) will be internally stable.

Proof. By Assumption 1, G is stable; by Lemma 1, $R(1+GR)^{-1} = TG^{-1}$ is stable; since G is strictly proper and R is proper, TG^{-1} is proper. Apply Lemma 11 of Vidyasagar (1985).

In the next section, conditions on the set $\{R\}$ are developed which guarantee the existence of solutions to Problem 2.

4. Arbitrarily small sensitivity

This problem includes simultaneous stabilization as well as robust performance. This is naturally a more difficult problem than that of simultaneous stabilization alone, so that further assumptions are required besides those made in Vidyasagar (1985).

4.1. The SISO case.

Lemma 2. Suppose that for all $R \in \{R\}$, $R \neq 0$, R has no zeros in the RHP, and all $R \in \{R\}$ have the same HFG sign; then $\{R\}$ will be an arbitrarily small sensitivity set. Moreover, for any given positive integer l and $0 < x < 1$, there exists a stable and minimum phase controller G with l pole excess over zeros, such that $|1+GR| > x$ or $|GR/(1+GR)| < 1/x$.

Remark 1. Clearly, x is related to the gain and phase margin of a SISO system; for example, $x = 0.5$ gives a 30° phase margin and 6db gain margin for $|1+GR| > x$.

Proof. A constructive proof limited to a fixed pole/zero excess in $\{R\}$ is given in Horowitz (1979, Appendix 1). This can be extended by induction to the non fixed pole/zero excess. In Wei and Yedavalli (1989, Theorem 5.5), there is a constructive proof to find a fixed controller for the simultaneous stabilization of sets, including the sets defined in Lemma 2. Using $|r(i\omega)| = 1/x$ in Corollary 4.3 in Wei and Yedavalli (1989), $|GR/(1+GR)| < 1/x$ is guaranteed. Their "procedure of synthesis" (Section 6.1) can be extended to guarantee that the open loop $|G(i\omega)R(i\omega)| > M$ if $\omega \leq \omega_0$ for any given M and ω_0 (select ϵ , small enough in Section 6.1); this will satisfy (2.2). A synthesis procedure for the design of F for a given G can be found in Horowitz (1979, Appendix 1).

One example of a set $\{R\}$ which violates these conditions is $\{R\} = \{s/(s-1), 1/(s-1)\}$. In this case, $s=0$ is a RHP zero of the first plant. It is not possible to simultaneously stabilize the two plants using a stable controller (G must have a pole in the origin in order to simultaneously stabilize this set). Another example is $\{R\} = \{1/(s+1), -1/(s+1)\}$; here the HFG sign is not fixed, so that it is possible to stabilize on condition that the loop transmission gain at $s=0$ is below 0dB. In the example $\{R\} = \{1/(s-1), -1/(s-1)\}$, the HFG sign is not fixed and no strictly proper stable controller exists which will provide simultaneous stabilization.

4.2. The MIMO case. Several notations are now introduced in order to reduce the complexity of the derivations. Let $q_k^0 = 1/P_{kk}^k$ where $g_i = 0$ is substituted in equation (3.3) for $i = 1, \dots, k-1$. Let R^k be the MTF R where rows and columns $1, \dots, k$ are deleted, and $R^n = 1$. That is, R^k is an $n-k$ by $n-k$ matrix, whose ij element is the $i+k, j+k$ element of R . Clearly, $R^0 = R$ and R^{n-1} is the nn element of R .

Lemma 3. Suppose the hypotheses of Lemma 1 are satisfied and $\det(R^{k-1})/\det(R^k) \neq 0$ for all $R \in \{R\}$ and $k = 1, \dots, n$. Under these conditions, there exists a finite minimum pole/zero excess for each g_k such that:

$$\text{HFG of } q_k = \text{HFG of } q_k^0 = \text{HFG of } \det(R^{k-1})/\det(R^k). \quad (4.1)$$

Proof. The right side of equation (4.1) is shown in Appendix Lemma A2. Substituting $m = i = j = k$ in equation (A6a) of the Appendix gives:

$$V_{kk}^k = P_{kk}^k = 1/q_k^k \quad (4.2)$$

Application of Lemma A1 of the Appendix to $R^{-1} = V_k$ makes it possible to calculate V_{kk}^k as follows: (1) Take the

inverse of V_k ; (2) Erase rows and columns $1, \dots, k-1$ of the result; (3) V_{kk}^A is the k th element of the inverse of the result. Since $V_k = R^{-1} + \text{diag}[g_1, \dots, g_{k-1}, 0, \dots, 0]$, V_{kk}^A has the following form:

$$\frac{1}{V_{kk}^A} = \frac{A + F_1(g_1, \dots, g_{k-1})}{B + F_2(g_1, \dots, g_{k-1})} \quad (4.3)$$

where F_1 and F_2 are sums of the multiplication of at least one of the g_i s by elements of R , so that A/B can be calculated by substituting $g_i = 0$ in equation (4.3) or in V_k . Since substituting $g_i = 0$ in V_k gives $V_k = R^{-1}$, by Lemmas A1 and A2 in the Appendix:

$$q_k = \frac{A}{B} \quad (4.4)$$

Since by the Lemma assumptions $A/B \neq 0$, one can calculate the minimum pole/zero excess required in each g_i for the pole/zero excess of A/B to be less than the pole/zero excess of F_1/F_2 . In Lemma 1, it is shown that P_{kk}^A is RRHP, hence:

$$\text{HFG of } q_k = \lim_{|s| \rightarrow \infty} \left\{ \frac{s^l}{V_{kk}^A} \right\} = \lim_{|s| \rightarrow \infty} \left\{ s^l \frac{A}{B} \right\} \text{ as } |s| \rightarrow \infty = \text{HFG of } q_k^0. \quad (4.5)$$

An example of Lemma 3 for the two-input-output case is now presented

$$V_1 = \text{inv}(R^{-1} + \text{diag}(g_1, 0)) \quad (4.6)$$

hence,

$$\frac{1}{P_{11}^A} = \frac{P_{22}P_{11} - P_{12}P_{21} + P_{22}g_1}{P_{22}}, \quad R^{-1} = [P_{ij}] \quad (4.7)$$

and the pole/zero excess required for g_1 is more than the pole/zero excess of $1/P_{22}$. The next theorem is our main result.

Theorem 2. Suppose that the hypotheses of Lemma 3 are satisfied, and that for $k = 1, \dots, n$ the HFG of $\det(R^k)/\det(R^A)$ is fixed for all $R \in \{R\}$, then $\{R\}$ is an arbitrarily small sensitivity set

Proof. This is shown by induction on the n steps of the Horowitz method for MIMO systems. The induction hypotheses for step k are: the designer can choose g_k and f_k , such that: (1) Inequalities (3.8-3.10) are satisfied; (2) R_k is stable with no RHP zeros, (3) $(g_k + P_{kk}^A)$ has no RHP zeros over all $R \in \{R\}$, and (4) g_k has sufficient pole/zero excess such that the HFG of q_k is the same as the HFG of q_k^0 (see Lemma 3)

First induction step, $k = 1$: Since $q_1 = \det(R)/\det(R^1)$, by hypotheses it follows that the HFG of q_1 is fixed, that $q_1 \neq 0$, and that q_1 has no RHP zeros. Thus the conditions of Lemma 2 are satisfied, hence there exists a g_1 such that the induction hypotheses for $k = 1$ are satisfied (a minimum of pole/zero excess in g_1 has to be selected).

Next induction step: By the induction hypotheses on step k and Lemma 3, the HFG of q_{k+1} is the same as the HFG of q_{k+1}^0 which by the theorem assumptions is fixed $\neq 0$, so also $q_{k+1} \neq 0$. By Lemma 1, q_{k+1} has no RHP zeros. Thus the conditions of Lemma 2 are satisfied for $k+1$, and the designer can choose g_{k+1} and f_{k+1} with sufficient pole/zero excess such that the induction hypotheses are satisfied.

When a given set $\{R\}$ does not satisfy the hypotheses of Theorem 2, it may still be an arbitrarily small sensitivity set. This is shown as follows.

Remark 2. If there exists an MTF G_1 such that $\{RG_1\}$ satisfies the hypotheses of Theorem 4.4, apply the design method on the set $\{RG_1\}$ in order to find a controller G_2 . The controller which solves the problem is G_1G_2 . Hence our main result is not restricted to a diagonal design of G .

5. Conclusions

If for all $R \in \{R\}$: (1) R^{-1} is RRHP and R is proper, and (2) the sign of the high frequency gain of the leading principal minor $(k-1)$ divided by the leading principal

minor k is fixed and not zero for $k = 1, \dots, n$, then the Horowitz design method for uncertain MIMO feedback systems can be used to find an internally stable bandwidth limited solution for any desired closed loop response tolerances and arbitrarily high disturbance rejection specifications. A key requirement for the controller elements is a minimum pole/zero excess. This paper assumes that $\{R\}$ is a finite set, due to modelling error $\{R\}$ can be an infinite set and it is possible to show examples of infinite sets $\{R\}$ for which the results are not true, even in the SISO case.

Acknowledgements—This research was supported by grant No. 86-00034 from the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel, and the National Science Foundation Grant ECS-8-608875 at the University of California.

References

- Barmish, B. and S. Zhicheng (1988). A simplified test for robust stability of delay systems. *Proc. 27th CDC*, pp. 92-97.
- Bode, H. W. (1945). *Networks Analysis and Feedback Amplifier Design*. Van Nostrand, New York.
- Gantmacher, F. R. (1960). *The Theory of Matrices*, Chelsea, New York.
- Garnell, P. and D. J. East (1977). *Guided Weapon Control Systems*. Pergamon, Oxford.
- Ghosh, K. (1986). Simultaneously partial pole placement: A new approach to multimode system design. *IEEE Trans. Aut. Control*, **AC-31**, 440-443.
- Horowitz, I. (1979). Quantitative synthesis of uncertain multiple input-output feedback systems. *Int. J. Control*, **30**, 81-106.
- Horowitz, I. (1982). Improved design technique for uncertain multiple-input-output feedback systems. *Int. J. Control*, **36**, 977-988.
- Paul, R. P. (1981). *Robot Manipulators. Mathematics, Programming and Control*. MIT Press, Cambridge, MA.
- Saber, A. (1985). Simultaneous stabilization with almost disturbance decoupling part 1: Uniform rank systems. *24th IEEE Conf.*, **1**, 306-308.
- Schmitendorf, W. E. and C. V. Hollot (1988). Simultaneous stabilization via linear state feedback control. *Proc. 27th CDC*, Austin, TX, pp. 1781-1786.
- Skogestad, S., M. Morari and J. Doyle (1988). Robust control of ill-conditioned plants. High-purity distillation. *IEEE Trans. Aut. Control*, **AC-33**, 1092-1105.
- Vidyasagar, M. (1985). *Control System Synthesis: A Factorization Approach*. MIT Press, Cambridge, MA.
- Wei, K. and R. K. Yedavalli (1989). Robust stabilizability for linear systems with both parameter variation and unstructured uncertainty. *IEEE Trans. Aut. Control*, **AC-34**.
- Wei, K. and B. Barmish (1988). An iterative design procedure for simultaneous stabilization of MIMO systems. *Automatica*, **24**, 643-652.
- Wie, B. and K. W. Byun (1989). New approach to attitude momentum control for the space station. *Guidance*, **7**, 714-722.
- Yaniv, O. and I. Horowitz (1986). A quantitative design method for MIMO linear feedback systems having uncertain plants. *Int. J. Control*, **46**, 401.

Appendix: Proofs of Lemmas

Let:

$R =$ an n by n MTF

$P = R^{-1}$

$R^k =$ The MTF R where rows and columns $1, \dots, k$ are deleted. That is, R^k is an $n-k$ by $n-k$ matrix, whose ij element is the $i+k, j+k$ element of R . In addition define $R'' = R$ and $R'' = 1$

$\Delta^k = \det(R^k)$

$\Delta_{ij}^k = ij$ cofactor of R^k

$\Delta_{ab,cd}^k = (-1)^{a+b+c+d} [\det \text{ of } R^k \text{ after rows } a, c \text{ and columns } b, d \text{ are deleted}]$

$[P_{ij}^k] = \Delta n - k + 1$ by n MTF defined recursively:

$$[P_{ij}^1] = R^{-1}, \quad P_{ij}^{k+1} = P_{ij}^k - \frac{P_{ik}^k P_{kj}^k}{P_{kk}^k};$$

$k = 1, \dots, n-1$ and $(i = k+1, \dots, n, j = 1, \dots, n)$. (A1)

Lemma A1. $\text{inv}(R^{k-1}) = [P_{ij}^k]$ for $k = 1, \dots, n$, and $i, j = k, \dots, n$.

Proof. By induction.

The case $k = 1$. For $k = 1$, $\text{inv}(R^0) = R^{-1} = [P_{ij}^1]$, hence the induction hypothesis is satisfied.

The general case. Using the above notations, the induction hypothesis is: $[P_{ij}^k] = \Delta_{ij}^{k-1} / \Delta^{k-1}$. The identity (Bode, 1945, p. 54):

$$\Delta_{ab} \Delta_{cd} - \Delta_{ad} \Delta_{cb} = \Delta \Delta_{ab, cd} \tag{A2}$$

is applied to R^{k-1} (whose cofactors are $\Delta^{k-1} P_{ij}^k$), which gives:

$$P_{ij}^k P_{kk}^k - P_{ik}^k P_{kj}^k = \frac{\Delta_{ij}^{k-1, kk}}{\Delta^{k-1}} - \frac{\Delta_{ij}^k}{\Delta^{k-1}}. \tag{A3}$$

Thus,

$$P_{ij}^{k+1} = P_{ij}^k - \frac{P_{ik}^k P_{kj}^k}{P_{kk}^k} = \frac{\Delta_{ij}^{k-1, kk}}{\Delta^{k-1} P_{kk}^k} - \frac{\Delta_{ij}^k}{\Delta^{k-1}}. \tag{A4}$$

The right side of (A4) proves the induction hypothesis for step $k+1$.

Lemma A2. Suppose R is an n by n MTF, then:

$$q_k^0 = \frac{\det(R^{k-1})}{\det(R^k)}. \tag{A5}$$

Proof. $q_k^0 = 1/P_{kk}^k$ is calculated recursively by (3.3) for $g_i = 0$ ($i = 1, \dots, k-1$), which is the same recursive equation (A1) for the plant R . Hence (A5) is an immediate consequence of Lemma A1

Lemma A3. Let R be a square n by n MTF and g_1, \dots, g_m be n TF. Define:

$P = [P_{ij}] = R^{-1}$
 $V_m = [P_{ij}] + \text{diag}[g_1, \dots, g_{m-1}, 0, \dots, 0]$
 P_{ij}^k = Defined in equation (3.3)
 V_{ij}^k = Defined in equation (A1), where V_m replaces R^{-1} and V_{ij}^k replaces P_{ij}^k . The index m is omitted for clarity.

Then for $k = 1, \dots, m$

$$V_{ij}^k = P_{ij}^k \text{ for } (i \neq j \text{ or } i \geq m) \text{ and } i = k, \dots, n \tag{A6a}$$

$$V_{ii}^k = P_{ii}^k + g_i \text{ for } i = k, \dots, m-1 \tag{A6b}$$

Proof. For a given m by induction on k . For $k = 1$ trivial, assume equations (A6) are true for $k = r$, then by definition:

$$V_{ij}^{r+1} = V_{ij}^r - \frac{V_{ir}^r V_{rj}^r}{V_{rr}^r} \tag{A7}$$

Substituting the induction hypothesis in equation (A7) gives:

$$V_{ij}^{r+1} = P_{ij}^r - \frac{P_{ir}^r P_{rj}^r}{P_{rr}^r + g_r} = P_{ij}^{r+1} \text{ for } (i \neq j \text{ or } i \geq m) \text{ and } i = r+1, \dots, n \tag{A8a}$$

$$V_{ii}^{r+1} = P_{ii}^r + g_i - \frac{P_{ir}^r P_{ri}^r}{P_{rr}^r + g_r} = P_{ii}^{r+1} + g_i \text{ for } i = r+1, \dots, m-1. \tag{A8b}$$

Approximation of Delay Systems by Fourier–Laguerre Series*

JONATHAN R. PARTINGTON†

Key Words—Approximation theory; convergence analysis; delays; error analysis; Fourier analysis; Laplace transforms; model reduction; system order reduction

Abstract—Error estimates are given for the approximation of stable recorded delay systems in the L_2 and H_2 norms, using two recently advocated techniques based on Laguerre series. In addition, some theoretical results on $L_2(0, \infty)$ approximation are derived.

1. Introduction

CONSIDER A stable continuous-time linear time-invariant system, (which will usually be infinite-dimensional), with impulse response $g(t)$ defined on $(0, \infty)$ and transfer function the Laplace transform of g , i.e.

$$G(s) = \int_0^\infty e^{-st} g(t) dt$$

We restrict our discussion to single-input single-output systems, for simplicity of notation, however our results extend to multiple inputs and outputs. The problem of approximating g (or G) in various norms is one which has been much studied in recent years. In particular we shall be interested in the following norms

$$\|g\|_2 = \left(\int_0^\infty |g(t)|^2 dt \right)^{1/2}$$

(the L_2 norm) and

$$\|G\|_2 = \sup \{ \|G(s)\| : \operatorname{Re} s < 0 \}$$

(the H_2 norm)

Note that we also have

$$\|g\|_2 = \left(\frac{1}{2\pi} \int_{-\infty}^\infty |G(i\omega)|^2 d\omega \right)^{1/2}$$

[see e.g. Partington (1988), Chapter 2], and we shall use the notation $\|G\|_2$ to denote $\|g\|_2$.

The H_2 norm of a system has been shown to be of interest for problems of control system synthesis (Francis, 1987), whereas the L_2 norm finds applications to questions of system identification (see e.g. Wahlberg and Ljung, 1986) as well as in LOG control; both norms have been considered from the point of view of model reduction (i.e. rational approximation) in a large number of papers.

In Section 2, we discuss two recently suggested techniques for approximating systems, (Gu *et al.*, 1989; Makila, 1990b), both based on Fourier–Laguerre series; these have advantages through being easier to calculate than more rapidly converging approximations (such as the Hankel-norm and truncated balanced realization methods), and it is

therefore helpful to have theoretical results indicating how fast they do converge.

In many ways the most common, as well as the simplest infinite-dimensional systems to analyse are those arising from delay differential equations (delay systems), and in particular the retarded delay systems (defined in Section 3). We give a detailed analysis of convergence rates of the two methods based on the concept of the *index* of a delay system, a notion which is of importance in other methods of approximation.

In Section 4 we give a discussion of a somewhat different norm, the time-domain L_∞ norm, and give some theoretical results on approximation in this norm, with special reference to delay systems.

Finally, Section 5 contains an example: we analyse one particular delay system in detail, and review all the approximation schemes for which error bounds are available.

2. Two techniques for approximating systems

We shall analyse two techniques for approximating systems by systems with a single (repeated) pole. Both have appeared recently in the literature and have applications to delay systems.

The first approximation technique that we shall discuss is given by Gu *et al.* (1989), and involves using the isometry between $H_2(\mathbb{C}_+)$ and the classical H_2 space on the disc, induced by the conformal map $z = Ms = (\lambda - s)/(\lambda + s)$, $s = M^{-1}z = \lambda(1 - z)/(1 + z)$; here λ is a fixed positive real number. In general we shall not be concerned with the actual value of λ , since the estimates hold whichever value we take; however, there is evidence (see e.g. Glover *et al.*, 1990c) that best results may be obtained by taking λ to be of the same order as n , the degree of the approximation.

We are therefore led to consider the following expansion of a transfer function $G(s)$:

$$G(s) = \sum_{k=0}^\infty c_k \left(\frac{\lambda - s}{\lambda + s} \right)^k, \quad (2.1)$$

where the (c_k) are constants (in fact Gu *et al.* consider the slightly more general matrix-valued case but this adds few extra difficulties).

Moreover, if we write $H(s) = \sqrt{2\lambda}/(\lambda + s)G(s)$ then we have

$$H(s) = \sum_{k=0}^\infty c_k \sqrt{2\lambda} \frac{(\lambda - s)^k}{(\lambda + s)^{k+1}},$$

so that we can identify the constants (c_k) as the coefficients in the Fourier expansion of $H(s)$ with respect to the orthonormal basis

$$(c_k(s)) = \sqrt{2\lambda} \frac{(\lambda - s)^k}{(\lambda + s)^{k+1}}$$

of the Hardy space $H_2(\mathbb{C}_+)$ using the norm $\|G\|_2$ above, induced from $L_2(0, \infty)$. (For simplicity of notation we suppress the dependence on λ .)

Let $L_k(t) = e^{-t}/k! \, d^k/dt^k (e^t)$ be the classical Laguerre polynomial [see e.g. Szegő (1939) for details]. Then the

* Received 9 January 1990; revised 24 July 1990; received in final form 22 August 1990. The original version of this paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor H. Kimura under the direction of Editor H. Kwakernaak.

† School of Mathematics, University of Leeds, Leeds LS2 9JT, U.K.

functions (c_k) are the Laplace transforms of the functions $\phi_k(t) = (-1)^k \sqrt{2\lambda} e^{-\lambda t} L_k(2\lambda t)$, which form an orthonormal basis in the space $L_2(0, \infty)$

A degree- n approximation to $G(s)$ is given by

$$G^n(s) = \sum_{k=0}^n c_k \left(\frac{\lambda - s}{\lambda + s} \right)^k \quad (2.2)$$

Clearly $H_n(\cdot, \cdot)$ convergence of the sequence (G^n) to G is equivalent to H_n convergence of the power series expansion of $G(Ms)$ on the disc. We cannot expect to obtain L_2 convergence of the series (G^n) , however, although we shall obtain L_2 convergence of the series $H_n(s) = \sqrt{2\lambda} G^n(s)/(\lambda + s)$ to H whenever $H(s)$ is in $H_2(\cdot, \cdot)$.

The second technique consists in truncating a Fourier expansion

$$G(s) = \sum_{k=0}^{\infty} d_k \sqrt{2\lambda} \frac{(\lambda - s)^k}{(\lambda + s)^{k+1}}$$

to obtain an approximation

$$G_n(s) = \sum_{k=0}^n d_k \sqrt{2\lambda} \frac{(\lambda - s)^k}{(\lambda + s)^{k+1}}$$

Again we suppress the dependence on λ in the notation. Note our notational convention that G_n and H_n are truncated Fourier series, whereas G^n is not.

The constants (c_k) and (d_k) are related as follows: we have that

$$G(M^{-1}z) = \sum_{k=0}^{\infty} c_k z^k = \sum_{k=0}^{\infty} d_k z^k \frac{(1+z)}{\sqrt{2\lambda}}$$

and hence $d_0/\sqrt{2\lambda} = c_0$ and $(d_k + d_{k-1})/\sqrt{2\lambda} = c_k$ for $k \geq 1$.

This technique was used in Glover *et al.* (1990c) with λ chosen to depend on n , and has been recently employed by Mäkilä in a number of papers (including Mäkilä, 1990a, b). Indeed Mäkilä (1990b) gave an explicit error estimate for this technique applied to simple delay systems of the form $G(s) = e^{-\tau s} R(s)$ with R rational and strictly proper and $\tau > 0$. Namely, it showed that if $r \geq 1$ is the relative degree of R (that is, the denominator degree is r more than the numerator degree), then

$$d_k = O(k^{-1/4 - r/2}) \quad (2.3)$$

(this estimate is tight) and so

$$c_k = O(k^{-1/4 - r/2}). \quad (2.4)$$

Hence we obtain

$$\|G - G_n\|_2 = O(n^{1/4 - r/2})$$

and can estimate $\|G - G_n\|_\infty$ as follows.

$$\begin{aligned} (G - G_n)(M^{-1}z) &= \sum_{k=n+1}^{\infty} d_k z^k \frac{(1+z)}{\sqrt{2\lambda}} \\ &= \sum_{k=n+1}^{\infty} (d_k + d_{k-1}) \frac{z^{k-1}}{\sqrt{2\lambda}} = d_n z^{n+1} / \sqrt{2\lambda}, \end{aligned}$$

and so

$$\begin{aligned} \|G - G_n\|_\infty &\leq \sum_{k=n+1}^{\infty} |c_k| = |d_n| / \sqrt{2\lambda} \\ &= O(n^{1/4 - r/2}). \end{aligned}$$

(This is a similar argument to the one given in Mäkilä, 1990b). Note that for any sequence $(c_k) = O(k^{-\alpha})$ and $\alpha > 1$, we have $\sum_{k=n+1}^{\infty} |c_k| = O(n^{1-\alpha})$, since

$$\sum_{k=n+1}^{\infty} k^{-\alpha} \leq \int_n^{\infty} x^{-\alpha} dx = n^{1-\alpha} / (\alpha - 1).$$

We shall use this estimation technique several times.

In Section 3 we shall extend these results to general retarded delay systems; however we remark now that with $G(s) = e^{-\tau s} R(s)$, and r as above, the optimal convergence rate achievable with any rational approximation of G is given

by

$$\|G - \tilde{G}\|_2 \geq A n^{1/2 - r} \quad (2.5)$$

and

$$\|G - \tilde{G}\|_\infty \geq B n^{-r} \quad (2.6)$$

for some constants A and B —see Glover *et al.* (1990b, c) for details—so that there is a trade-off here between ease of calculation and accuracy of approximation.

3. Decomposition of retarded systems and error bounds

The most general SISO retarded delay system has the transfer function

$$G(s) = h_2(s)/h_1(s), \quad (3.1)$$

where

$$h_1(s) = \sum_{i=0}^{n_1} p_i(s) e^{-\gamma_i s}, \quad (3.2)$$

and

$$h_2(s) = \sum_{i=0}^{n_2} q_i(s) e^{-\beta_i s}, \quad (3.3)$$

with $0 = \gamma_0 < \gamma_1 < \dots < \gamma_{n_1}$, $0 = \beta_0 < \dots < \beta_{n_2}$, the p_i being polynomials of degree δ_i , and $\delta_i < \delta_0$ for $i \neq 0$, and the q_i polynomials of degree $d_i < \delta_0$ for each i . Such functions were analysed in Bellman and Cooke (1963), where it was shown that they have only finitely many poles in any right half plane. They have since been considered from the viewpoint of approximation in Zwart *et al.* (1988), Partington *et al.* (1988), Glover *et al.* (1990a, b, c) and Partington and Glover (1990), and we begin by summarising some of what is known about such systems.

It is convenient to define the *index* of a delay system, $I(G)$ to be the unique value $r \geq 1$ such that one may write

$$G(s) = R(s) + \sum_{i=1}^P a_i e^{-\alpha_i s} / (s+1)^i + O(s^{-1}), \quad (3.4)$$

with R rational, $P \geq 1$, the (a_i) nonzero constants, and the (α_i) positive coefficients.

Thus for example $e^{-2s}/(s+e^{-s}) = e^{-2s}/(s+1) + O(s^{-2})$, so it has index 1, whereas $1/(s+e^{-s}) = 1/(s+1) + (1 - e^{-s})/((s+1)(s+e^{-s})) = 1/(s+1) + 1/(s+1)^2 + e^{-s}/(s+1)^2 + O(s^{-3})$, and has index 2. For the case $G(s) = e^{-\tau s} R(s)$ with R rational, (i.e. $n_1 = 0$), the index $I(G)$ is just the relative degree of R .

An equivalent formulation may be derived from Glover *et al.* (1990a, Section 5 and 1990b, Section 5), where it is shown that $I(G)$ is such that the Hankel singular values of G and the minimum H_∞ error in approximating G by a system of degree n are bounded above and below by multiples of $n^{-I(G)}$. Similarly in Glover *et al.* (1990c) it is shown that the optimal L_2 approximation error is of order $n^{1/2 - I(G)}$. The key to these discussions is that the dominant term which influences the approximation rate is the term $\sum_{i=1}^P a_i e^{-\alpha_i s} / (s+1)^i$ occurring in (3.4). In the time domain this corresponds to analysing the impulse response g corresponding to G for which $g^{(k-1)}$ is the first discontinuous derivative.

It is possible to approximate unstable delay systems by writing them as $G(s) = G_{\text{stable}}(s) + K(s)$, where K is rational, strictly proper and totally unstable, and then approximating G_{stable} , but other approaches are often more satisfactory, e.g. the one based on coprime factorizations (Partington and Glover, 1990). Clearly the index of an unstable system is the same as the index of its stable part, but we shall assume in what follows that G is stable. As in Section 2, we write (c_k) to denote the Fourier coefficients of the function $H(s) = \sqrt{2\lambda} G(s)/(\lambda + s)$ and we write (d_k) for the Fourier coefficients of $G(s)$ itself.

Our first result is a consequence of the results of Mäkilä and the decomposition technique of Glover *et al.* (1990a, b, c).

Lemma 1. Let G be a stable retarded delay system of index

$r \geq 1$. Then the coefficients (c_k) and (d_k) satisfy

$$c_k = O(k^{-1/4-r/2})$$

and

$$d_k = O(k^{-1/4-r/2}) \quad \text{as } k \rightarrow \infty. \quad (3.5)$$

Both these estimates are tight in general.

Proof. Since $H(s)$ is clearly a retarded delay system of index $(r+1)$ it is sufficient to prove the result for d_k . We introduce the notation $T_r(s) = \sum_{i=0}^r a_i e^{-i/s} / (s+1)^r$ [cf. (3.4)], and observe that it is possible to iterate the decomposition method and write, for any $N \geq r$,

$$G(s) = R_N(s) + T_r(s) + T_{r+1}(s) + \dots + T_N(s) + S_N(s),$$

where R_N is rational, each T_i is equal to a sum of delay terms divided by $(s+1)^i$, and $S_N(s) = O(s^{-N-1})$.

Now it is easy to verify that $d_k(R_N) \rightarrow 0$ exponentially for any stable rational function, (e.g. by transforming to the disc and examining the power series); moreover each T_i has the property that $d_k(T_i) = O(k^{-1/4-i/2})$, by (2.3). Finally, for sufficiently large N the coefficients $d_k(S_N)$ are $o(k^{-1/4-r/2})$, by the results in Section 6 of Glover *et al.* (1990c), specifically since by Lemma 6.6 of that paper S_N has an impulse response with $(N-1)$ continuous derivatives, and exponential decay at ∞ ; integration by parts shows that the Fourier-Laguerre coefficients of a function f of exponential decay will decrease faster than any given power of k if f is sufficiently smooth.

It follows that $d_k(G) = O(k^{-1/4-r/2})$, as required. To see that the estimate is tight, note that it is sufficient to verify this for $d_k(T_r)$, since all the other terms are dominated by this. Apart from exceptional cases in which coefficients cancel (this cannot happen for $P=1$ and may be impossible in general), this will be the case.

Corollary 1. Let $G(s)$ be a stable retarded delay system of index r . Then the following bounds hold

$$\|G - G^n\|_2 = O(n^{1/4-r/2}) \quad (3.6)$$

$$\|G - G_n\|_2 = O(n^{1/4-r/2}) \quad (3.7)$$

$$\|G - G_n\|_2 = O(n^{1/4-r/2}) \quad (3.8)$$

Proof. Since $\|((\lambda-s)/(\lambda+s))^k\|_2 = 1$, for all k , it follows that $\|G - G^n\|_2 \leq \sum_{k=n+1}^{\infty} |c_k|$ and the first result follows because $c_k = O(k^{-1/4-r/2})$. Alternatively, for $r \geq 2$ one can use the error estimate given in Gu *et al.* (1989, Theorem 2.17), namely

$$\|G - G^n\|_2 \leq \sqrt{2\lambda/n} \left(\sum_{k=n+1}^{\infty} k^2 c_k^2 \right)^{1/2},$$

which gives a bound of $O(n^{-1/2} (\sum_{k=n+1}^{\infty} k^{1-2r})^{1/2})$, which is $O(n^{1/4-r/2})$.

The other two estimates follow from (3.5) since the arguments given in Section 2 go through to this more general

4. Approximation in the $L_2(0, \infty)$ norm

It is of interest in some applications to consider uniform approximation of the impulse response of a linear system, and accordingly we write

$$\|g\|_{\infty} = \text{ess. sup}_{0 \leq t < \infty} |g(t)|$$

for the norm in $L_2(0, \infty)$. It is our first concern to derive a lower bound for the error in approximating a linear system in this norm; we then compare the various techniques available.

Although the $L_2(0, \infty)$ norm is an operator norm, being the norm of the Hankel operator

$$(Gu)(t) := \int_0^{\infty} g(t+\tau)u(\tau) d\tau$$

as a map from $L_2(0, \infty)$ to $L_2(0, \infty)$, little appears to be little known about the approximation numbers of operators between these spaces, and we proceed differently. For g in

$L_2(0, \infty)$ we write $E_k(g)$ to denote the minimum L_2 error of any degree- k rational approximation to g .

Theorem 1. Let $g(t)$ be a function in $L_2(0, \infty)$. Then, for any $\hat{g}(t)$ in $L_2(0, \infty)$ whose Laplace transform is rational of degree n , we have

$$\|g - \hat{g}\|_{\infty} \geq \sup (E_n(g(t)e^{-\mu t})/\sqrt{2\mu}; \mu > 0).$$

Proof. It is a standard result that $\|f\|_{\infty} = \sup \{ \|fh\|_2 / \|h\|_2 : h \in L_2, h \neq 0 \}$. Hence

$$\|g - \hat{g}\|_{\infty} \geq \frac{\|(g - \hat{g})e^{-\mu t}\|_2}{\|e^{-\mu t}\|_2} = \|(g - \hat{g})e^{-\mu t}\|_2 / \sqrt{2\mu}.$$

Since $ge^{-\mu t}$ also has a rational Laplace transform of degree n , the result now follows.

Corollary 2. Let $G(s)$ be a retarded delay system of index r with impulse response $g(t)$. Then there exists a constant $A > 0$ such that for any degree- n approximant \hat{G} with impulse response \hat{g} we have

$$\|g - \hat{g}\|_{\infty} \geq \begin{cases} An^{-1} & \text{if } r \geq 2; \\ A & \text{if } r = 1. \end{cases}$$

Proof. This follows from Theorem 1 (we may take $\mu = 1$) and the known L_2 result (2.5), given in Glover *et al.* (1990c) for $r \geq 2$. In the case $r = 1$ the impulse response g is discontinuous and so it cannot be approximated uniformly by continuous impulse responses. \square

This inequality is not tight, and there is some evidence to suggest that the "correct" inequality should be $\|g - \hat{g}\|_{\infty} \geq An^{-1}$ for all values of r .

As regards achievable $L_2(0, \infty)$ errors, little is known. One possible technique is to approximate $\hat{g}(t)$ in the L_1 norm (in the case $r \geq 2$, naturally), since $\|g - \hat{g}\|_{\infty} \leq \|g - \hat{g}\|_1$, provided that $g(\infty) = \hat{g}(\infty) = 0$. The methods of Glover and Partington (1987) can be used here and it can be shown that an achievable error bound is $O((\log n/n)^{1/2})$.

Alternatively one can use the truncated series method of Sections 2 and 3, and, since $\|\phi_k\|_2 \leq 1$ for all k (Szegő, 1939) we obtain the error bound

$$\|g - g_n\|_2 \leq \sum_{k=n+1}^{\infty} |d_k| = O(n^{1/4-r/2}).$$

5. Example

Consider the linear system with transfer function

$$G(s) = \frac{1}{s+1-\exp(-2-s)}.$$

This retarded delay system has been considered in several places (Partington *et al.*, 1988; Gu *et al.*, 1989; Glover *et al.*, 1990b) for the purposes of assessing model reduction schemes and we collect together various results on this system. We begin with H_2 bounds.

The decomposition $G(s) = 1/(s+1) + e^{-2-s}/(s+1)^2 + O((s+1)^{-3})$ shows that the index of G is 2, and it also follows that the Hankel singular values satisfy

$$n^2 \sigma_n(G) \rightarrow \frac{1}{e^2 \pi^2} \quad \text{as } n \rightarrow \infty,$$

(cf. Glover *et al.*, 1990b), these provide an absolute lower bound for the approximation error by any scheme.

In Partington *et al.* (1988), two approximation schemes were given: the partial fraction scheme, with error $O(\log n/n)$, and the Padé-like scheme of approximating e^{-t} in the above expression by $((2n-s)/(2n+s))^n$: this has error $O(n^{-4/3})$. The best known theoretical bounds for Hankel-norm and truncated balanced realization approximations (Glover, 1984; Glover *et al.*, 1988) yield errors of $O(n^{-1})$ though in practice these methods appear to perform better than the currently known bounds would suggest. The techniques given in Glover *et al.* (1990b) (based on Padé approximation) achieve errors of $O(n^{-2})$ which is the optimal rate. Next, the technique given in Gu *et al.* (1989),

analysed in Sections 2 and 3, achieve errors of $O(n^{-3/4})$ by (3.6), since the index of G is 2, and finally the technique given in Mäkilä (1990b), also analysed above, also achieves errors of $O(n^{-3/4})$, by (3.7).

Fewer results are available for the L_2 norm. The theoretical (and achievable) lower bound, given in Glover *et al.* (1990c), is $O(n^{-1/2})$, since the singular values of the scaled Hankel operator are bounded above and below by multiples of $n^{-1/2}$.

The partial fraction approach (Partington, 1990) yields an error of $O(\log n/\sqrt{n})$. There is no error bound available for the optimal Hankel-norm approximation, and the error bound for the truncated balanced realization (Glover *et al.*, 1988) is not explicit. Finally, the truncated Laguerre series gives an error bound of $O(n^{-1/4})$, by (3.8).

6. Conclusions

Fourier-Laguerre series do not achieve the optimal convergence rate for model reduction of delay systems in general but they are usually easily computable and (with λ chosen appropriately) can provide a reasonably efficient approximation technique. Numerical results (see e.g. Glader *et al.*, 1990) suggest that for lower order approximations several techniques perform satisfactorily, so that each may be useful in certain circumstances. For higher order approximations (e.g. $n > 5$) the asymptotic behaviour starts to become more relevant and it is the case that a satisfactory approach in practice is to truncate a series to obtain a higher order approximation, and then perform a balanced or Hankel-norm reduction of the (large) finite-dimensional approximant.

References

- Bellman, R. and K. L. Cooke (1963) *Differential-Difference Equations*. Academic Press, New York.
- Francis, B. A. (1987). *A Course in H_∞ Control Theory*. Springer, Berlin.
- Glader, C., G. Högnäs, P. M. Mäkilä and H. T. Toivonen (1990). Approximation of delay systems—A case study. *Int. J. Control* (to appear).
- Glover, K. (1984). All optimal Hankel-norm approximations of linear multivariable systems and their L_∞ error bounds. *Int. J. Control*, **39**, 1115–1193.
- Glover, K., R. F. Curtain and J. R. Partington (1988). Realisation and approximation of linear infinite dimensional systems with error bounds. *SIAM J. Control*, **26**, 863–898.
- Glover, K., J. Lam and J. R. Partington (1986). Balanced realisation and Hankel-norm approximation of systems involving delays. *Proc. IEEE Conf. on Decision and Control*, Athens, pp. 1810–1815.
- Glover, K., J. Lam and J. R. Partington (1990a). Rational approximation of a class of infinite-dimensional systems I: Singular values of Hankel operators. *Math. Control Circ. Syst.*, **3**, 325–344.
- Glover, K., J. Lam and J. R. Partington (1990b). Rational approximation of a class of infinite-dimensional systems II: Optimal convergence rates of L_∞ approximants. *Math. Control Circ. Syst.* (to appear).
- Glover, K., J. Lam and J. R. Partington (1990c). Rational approximation of a class of infinite-dimensional systems: The L_2 case. *J. Approx. Theory* (to appear).
- Glover, K. and J. R. Partington (1987). Bounds on the achievable accuracy in model reduction. In *Modelling Robustness and Sensitivity Reduction in Control Systems*, NATO ASI series F, 95–118.
- Gu, G., P. P. Khargonekar and E. B. Lee (1989). Approximation of infinite dimensional systems. *IEEE Trans. Aut. Control*, **34**, 610–618.
- Mäkilä, P. M. (1990a). Approximation of stable systems by Laguerre filters. *Automatica*, **26**, 333–345.
- Mäkilä, P. M. (1990b). Laguerre series approximation of infinite dimensional systems. *Automatica*, **26**, 985–995.
- Partington, J. R. (1988). *An Introduction to Hankel Operators*. Cambridge University Press, Cambridge.
- Partington, J. R. (1990). The L_2 approximation of infinite-dimensional systems. *Proc. 5th I.M.A. Control Conf.* (to appear).
- Partington, J. R. and K. Glover (1990). Robust stabilization of delay systems by approximation of coprime factors. *Syst. Control Lett.*, **14**, 325–331.
- Partington, J. R., K. Glover, H. J. Zwart and R. F. Curtain (1988). L_∞ approximation and nuclearity of delay systems. *Syst. Control Lett.*, **10**, 59–65.
- Szegő, G. (1939). *Orthogonal Polynomials*. American Mathematical Society, New York.
- Wahlberg, B. and L. Ljung (1986). Design variables for bias distribution in transfer function estimation. *IEEE Trans. Aut. Control*, **31**, 131–144.
- Zwart, H. J., R. F. Curtain, J. R. Partington and K. Glover (1988). Partial fraction expansions for delay systems. *Syst. Control Lett.*, **10**, 235–244.

Brief Paper

The Diagonalizability of Quadratic Functions and the Arbitrariness of Shadow Prices*

BERÇ RUSTEM†

Key Words—Nonlinear programming; multiobjective optimization; quadratic programming; decision theory

Abstract—The possibility of tailoring the quadratic objective function to generate optimal policies which are acceptable to the policy maker is explored with two alternative algorithms. One of these is for objective functions with diagonal Hessians and uses updates of the desired values. The second algorithm proceeds by updating nondiagonal Hessians, while keeping the desired values fixed. The equivalence of both algorithms are established.

The arbitrariness of shadow prices are established for all quadratic equality constrained optimization problems. Starting initially with linear constraints, it is shown that either one of the algorithms can be used to alter the shadow prices, by arbitrarily specified amounts, without altering the primal optimal solution. In one step, the algorithms are able to change the shadow prices, while keeping the optimal policy fixed. It is shown that this is also true for nonlinear problems.

1. Introduction

WE CONSIDER two related issues concerning the quadratic objective function. The first is a method for the specification of the quadratic objective with a diagonal Hessian. The purpose of the method is to tailor the objective function such that the optimally generated solution of the decision problem is also politically acceptable to the decision maker. We study the equivalent of the method for nondiagonal Hessians. The method is then used to address the second problem which is the arbitrariness of the shadow prices, or Lagrange multipliers. The possibility of formulating optimization problems conveniently by using an appropriate Lagrangian is not new (see Geoffrion, 1971; Ponstein, 1983). However, in the present study we are concerned with the converse problem in which, while keeping fixed the (primal) optimal solution of the decision problem, we consider the possibility of altering, as required, the vector of (dual) shadow price, or Lagrange multiplier, values.

Consider the linear-quadratic optimal decision problem

$$\min \{ (x - x^d, D(x - x^d)) \mid N^T x = b \} \quad (1)$$

where $x \in E^n$, $b \in E^m$ is a fixed vector, D is a diagonal matrix

with non-negative elements, x^d is the desired or bliss value of x or the unconstrained optimum of the quadratic objective. The columns of matrix $N \in E^{n \times m}$ are assumed to be linearly independent. The diagonal weighting matrix in (1) occurs in economic decision making and in general in linear quadratic optimal control problems. In the latter case, (1) can be regarded as the transcription of the dynamic problem into static form (see e.g. Polak, 1971). We can state the two problems addressed below.

Problem 1. Let $\Omega \subseteq E^n$ denote the set of policies acceptable to the decision maker. We assume that there does not exist an analytical characterization of Ω and that Ω exists only in the mind of the decision maker. If an explicit characterization of Ω was possible, it could be used to augment the constraints of (1) and the resultant problem could easily be solved. An analytical characterization is assumed to be inherently difficult, or impossible, and sometimes also politically undesirable from the decision maker's point of view. The problem is to tailor an objective function for which the optimal solution of (1), x^* , also satisfies $x^* \in \Omega$.

We discuss two types of solutions to Problem 1. The first involves modifications to x^d and keeps D fixed as a diagonal matrix. The second involves modifications to D and generates a nondiagonal weighting matrix while keeping x^d fixed. We establish the equivalence of both methods.

Problem 2. Suppose that Problem 1 has been solved. Hence we have an objective function and $x^* \in \Omega$. The Lagrange multipliers of (1) signify shadow prices. Suppose that the shadow price vector, corresponding to x^* , is not acceptable. The problem is to alter the shadow prices to suit the decision maker's requirements while keeping the original x^* fixed.

By solving Problem 2, we ensure that both primal and dual variables satisfy the decision maker. It must be noted that Problem 2 is not the dual of Problem 1. In Problem 2, the primal solution, x^* , is fixed and the dual solution is altered, while in Problem 1 there is no question of fixing either the dual or the primal variables. We discuss two solutions to Problem 2. The first involves an alteration to x^d , while keeping D fixed, and the second involves the alteration of D to a general symmetric matrix, while keeping x^d fixed. The shadow prices can be assigned arbitrary desirable values without changing the optimal solution. Thus, the shadow prices of constrained optimization problems with quadratic objective functions are shown to be arbitrary. This can have serious implications in the use and interpretation of shadow prices (see e.g. UNIDO, 1971; Dreze, 1982).

The desirability of a diagonal D is due to computational reasons as well as the interpretation of the problem and the optimal solution. Nondiagonal weighting matrices can be diagonalized by appropriate transformations of the variables. However, such transformations lead to computational complications, particularly in relation to the constraints. A more important aspect is the difficulty of assigning an interpretation to the off-diagonal elements of a general symmetric weighting matrix. These elements are usually understood to represent trade-offs between achieving alternative objectives. However, they are difficult to assign an interpretation as distinct from the diagonal elements,

* Received 21 May 1989; revised 24 December 1989; revised September 1990, received in final form 6 September 1990. The original version of this paper was presented at the IFAC/SEDC/IFORS/IFIP Symposium on *Dynamic Modelling and Control of National Economies* which was held in Edinburgh, Scotland, U.K. during June, 1989. The published proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Associate Editor T. Başar under the direction of Editor A. P. Sage.

† Department of Computing, Programme of Research into Optimal Policy Evaluation, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London SW7 2BZ, U.K.

which represent the relative importance of each objective. Also, for linear quadratic dynamic problems, solved via dynamic programming, maintaining the block diagonality of D , in terms of the time structure, is desirable at least for the physical interpretation of the resultant optimal linear feedback laws. This contrasts with the case when nondiagonal weighting matrices have to be factorized and the original variables have to be transformed in order to obtain feedback laws which may only apply to the transformed variables.

In most economic decision making problems under conflicting objectives, the precise value of the weighting matrix is not known. Even a judicious choice of the weighting matrix need not necessarily yield a solution of problem (1) that is in the set of acceptable policies, Ω . In Section 2, we discuss an iterative algorithm, involving interactions with the decision maker, to specify a diagonal quadratic objective function that generates a solution of (1) which is *acceptable* to the decision maker. The algorithm involves the modification of the desired, or bliss, values of the objective function. In Section 3, the method is shown to be equivalent to an alternative method involving the respecification of the general non-diagonal symmetric matrix (Rustem *et al.*, 1978; Rustem and Velupillai, 1988a). In the latter method, the desired values remain fixed but the weighting matrix is altered and does not maintain a diagonal structure. This correspondence is helpful computationally as diagonal matrices are simpler to compute. Conversely, the correspondence provides the nondiagonal weighting matrix that would generate the same optimal solution as the one obtained by modifying the desired values. In addition, the correspondence can be used to discuss the complexity and polynomial time termination property of the algorithms by invoking the results in Rustem (1990) and Rustem and Velupillai (1988b). In Section 4, we discuss the use of both algorithms in solving Problem 2.

2. Specifying diagonal quadratic objective functions

We consider a solution to Problem 1. An alternative solution and the equivalence of both solutions are discussed in the next section. Let (1) be solved for a given weighting matrix D and a given initial vector of desired values x_0^d . The solution is denoted by

$$x_0 \approx \arg \min \{ (x - x_0^d, D(x - x_0^d)) \mid N^1 x = b \}.$$

The solution is presented to the decision maker who is required to respond by either declaring that $x_0 \in \Omega$, or if $x_0 \notin \Omega$, the decision maker is required to specify the modified form of x_0 that is in Ω .

The decision problem we wish to consider can now be formulated as the computation of the policy, optimally determined via (1), but which also is acceptable to the policy maker and is hence in the set Ω . We assume that $\{x \mid N^1 x = b\} \cap \Omega \neq \emptyset$. The decision maker's preferred alternative to x_0 is denoted by x_p . By definition, we have $x_p \in \Omega$ but not necessarily $x_p \in \{x \mid N^1 x = b\} \cap \Omega$. In case the latter is true, such a preferred alternative would conceptually solve the decision problem. Let $\delta_0 = x_p - x_0$, where δ_0 is the correction vector that needs to be added to x_0 in order to ensure that $x_0 + \delta_0 \in \Omega$. Using δ_0 , we can revise the desired or bliss value as $x_1^d = x_0^d + \alpha_0 \delta_0$ where $\alpha_0 \geq 0$ is a scalar. Using this new desired value, problem (1) is solved once again to yield a new optimal solution, x_1 . This solution is shown below to have desirable characteristics. However, as there is no guarantee that $x_1 \in \Omega$, the above procedure may need to be repeated. The resulting algorithm is summarized below. We discuss the complexity and termination properties of the algorithm in Rustem (1990) and Rustem and Velupillai (1988b).

Algorithm. Updates of desired values with a fixed diagonal weighting matrix.

Step 0: Given D , x_0^d , the sequence $\{\alpha_k\}$ and the constraints, set $k = 0$.

Step 1: Compute the solution of the optimization problem

$$x_k \approx \arg \min \{ (x - x_k^d, D(x - x_k^d)) \mid N^1 x = b \}. \quad (2)$$

Step 2: Interact with the decision maker. If $x_k \in \Omega$, stop. Otherwise, the decision maker is required to specify the preferred value x_p , and hence,

$$\delta_k = x_p - x_k. \quad (3)$$

Step 3: Update the desired values

$$x_{k+1}^d = x_k^d + \alpha_k \delta_k. \quad (4)$$

set $k = k + 1$ and go to Step 1.

The relationship of x_{k+1} , x_k and α_k , δ_k is summarized in the following results. The choice of the objective function may also be based on criteria other than $x_k \in \Omega$. For example, in the linear stochastic dynamical systems, the weighting matrices might be chosen to yield a stable minimum variance controller (see Engwerda and Otter, 1989). Also, the possibility of altering the desired values in order to change the solution has been considered previously [see e.g. Hughes-Hallet and Rees (1983)]. However, in the present study we consider the theoretical and equivalence properties of the approach adopted in this section to respecification, discussed in Section 3, of nondiagonal weighting matrices while keeping x^d fixed. The equivalence also provides the key to the complexity and convergence of the policy design process. In addition, as it is possible to decompose a symmetric matrix into a sequence of rank one updates (see Fiacco and McCormick, 1968), the method in this section and the equivalence result allow the possibility of expressing nondiagonal weighting matrices in terms of diagonalized objective functions.

An alternative characterization of Problem 1 would be in terms of finding a solution that simply satisfies the constraints and Ω , without any optimality requirement. One difficulty of this approach is that Ω needs to be explicitly specified. Let us assume that this was possible and that Ω was characterized by the intersection of a finite number of linear inequalities. It can then be shown that both the above algorithm and the algorithm in the next section are related to Khachian's (1979, 1980) algorithm for computing a feasible solution to a system of linear inequalities. In this framework, the vector δ is determined by one of the (linear) constraints bounding the set Ω . In particular, δ is related to the gradient of a constraint characterizing Ω , violated at x_k (Rustem and Velupillai, 1988b; Rustem, 1990). By invoking Khachian's result, we can obtain a polynomial time complexity for both algorithms. The added advantage of the two algorithms in this paper is that they relate each iteration with an objective function and optimality that is useful to the decision maker. Since the set of acceptable policies, Ω , does not exist anywhere except in the mind of the decision maker, the interpretation and specification of δ is aided by the optimality, at each iteration, of the quadratic objective function.

Proposition 1. Assume that D is positive semi-definite, the optimal solution and Lagrange multipliers can be written as

$$x_k + \alpha_k Z(Z^1 D Z)^{-1} Z^1 D \delta_k \quad (5)$$

where $Z \in E^{n \times (n-m)}$ is an orthogonal matrix* such that $Z^1 N = 0$, and

$$\lambda_{k+1} = \lambda_k - \alpha_k (N^1 N)^{-1} N^1 D [I - Z(Z^1 D Z)^{-1} Z^1 D] \delta_k. \quad (6)$$

Proof. To establish (5), we note that $x_{k+1} - x_k \in \{x \mid N^1 x = 0\}$ and, as $N^1 Z = 0$, any such vector can be written as a linear combination of the columns of Z . Thus, we have $x_{k+1} - x_k = Z w$, for some vector $w \in E^{n-m}$, and from the first order optimality condition of (2) we can write

$$Z^1 D [Z w + x_k - x_k^d - \alpha_k \delta_k] = Z^1 N \lambda_{k+1} = 0$$

* The choice of the orthogonal matrix Z is discussed further in Rustem and Velupillai (1988a). The numerically stable way of generating Z is by considering the QR decomposition of N . The matrix Z is given by the last $n-m$ columns of the matrix Q of this decomposition (see Gill *et al.*, 1981).

and thus

$$x_{k+1} - x_k = Zw = -Z(Z^T DZ)^{-1} Z^T D[x_k - x_k^d - \alpha_k \delta_k]. \quad (7)$$

From the optimality condition at iteration k , we have $Z^T D[x_k - x_k^d] = Z^T N \lambda_k = 0$. Thus, expression (5) follows from (7).

For (6), we use the optimality condition at $k+1$ to yield

$$\lambda_{k+1} = (N^T N)^{-1} N^T D[x_{k+1} - x_k + x_k - x_k^d - \alpha_k \delta_k]. \quad (8)$$

Using (5) and the optimality condition at k , $D(x_k - x_k^d) = N \lambda_k$, leads to expression (6). \square

3. The diagonal version of non-diagonal quadratic forms

We consider the equivalence of the algorithm in Section 2 with a method generating non-diagonal weighting matrices, discussed in Rustem *et al.* (1978) and Rustem and Velupillai (1988a). The complexity of the former is discussed in Rustem (1990) and Rustem and Velupillai (1988b), by exploiting this equivalence. The following algorithm uses the same δ_k as in the algorithm in Section 2. It keeps the desired values fixed but updates the weighting matrix of the quadratic optimization problem.

Algorithm. Fixed desired values and nondiagonal weighting matrix.

Step 0: Given a positive semi-definite weighting matrix Q_0 , the sequence μ_k , the desired values x^d and the constraints, set $k = 0$.

Step 1: Compute the solution of the optimization problem

$$x_k = \arg \min \{ (x - x^d, Q_k(x - x^d)) \mid N^T x = b \}.$$

Step 2: Interact with the decision maker. If $x_k \in \Omega$, stop. Otherwise, the decision maker is required to specify the preferred value x_p , and hence, $\delta_k = x_p - x_k$.

Step 3: Update the weighting matrix

$$Q_{k+1} = Q_k + \mu_k \frac{Q_k \delta_k \delta_k^T Q_k}{(\delta_k, Q_k \delta_k)}. \quad (9)$$

Set $k = k + 1$ and go to Step 1.

The matrix Q_k compared in the algorithm above is in general nondiagonal. It is shown below that starting with an initial diagonal matrix, the above algorithm and the algorithm in Section 1 are equivalent. At each stage, the nondiagonal version above has a constant diagonal equivalent in the algorithm of Section 2. The equivalent results to Proposition 1 related to the above algorithm are discussed in Rustem and Velupillai (1988a). We summarize these results. When Q_k is positive semi-definite, each subsequent iterate of the above algorithm is given by

$$x_{k+1} = x_k + \hat{\alpha}_k Z(Z^T Q_k Z)^{-1} Z^T Q_k \delta_k \quad (10a)$$

$$\lambda_{k+1} = \lambda_k - \hat{\alpha}_k (N^T N)^{-1} N^T Q_k (I - Z(Z^T Q_k Z)^{-1} Z^T Q_k) \delta_k \quad (10b)$$

$$\hat{\alpha}_k = - \frac{\mu_k (\delta_k, Q_k (x_k - x_k^d))}{(\delta_k, Q_k \delta_k) + \mu_k (Q_k \delta_k, Z(Z^T Q_k Z)^{-1} Z^T Q_k \delta_k)} \quad (10c)$$

(see Rustem and Velupillai, 1988a; Theorems 1 and 2, and Lemma 2).

We now show the equivalence of the solution of the two quadratic optimization problems: one with the diagonal Hessian fixed and only the desired or bliss values modified, and the other with the desired values fixed and only the diagonal matrix modified to a nondiagonal form.

Proposition 2. Let Q_k be nonsingular. Then, there exist μ_k , α_k and $\hat{\alpha}_k$ such that for

$$x_k^d + \alpha_k \delta_k: Q_{k+1} = Q_k + \mu_k \frac{Q_k \delta_k \delta_k^T Q_k}{(\delta_k, Q_k \delta_k)} \quad (11)$$

we have

$$\begin{aligned} \arg \min \{ (x - x_{k+1}^d, Q_k(x - x_{k+1}^d)) \mid N^T x = b \} \\ = \arg \min \{ (x - x_k^d, Q_{k+1}(x - x_k^d)) \mid N^T x = b \}. \end{aligned} \quad (12)$$

Moreover, we have $\hat{\alpha}_k = \alpha_k$. If α_k , μ_k are restricted such that α_k , $\mu_k \geq 0$, then the choice of δ_k is restricted by the inequality $(\delta_k, Q_k(x_k - x_k^d)) \leq 0$.

Proof.* Consider the optimality conditions of the minimization problems on both sides of (12). With the solution and the Lagrange multipliers denoted respectively by x_{k+1} , λ_{k+1} , the left side yields

$$Q_k(x_{k+1} - x_{k+1}^d) - N \lambda_{k+1} = 0.$$

The right side yields

$$\left[Q_k + \mu_k \frac{Q_k \delta_k \delta_k^T Q_k}{(\delta_k, Q_k \delta_k)} \right] (x_{k+1} - x_k^d) - N \lambda_{k+1} = 0.$$

Equating both optimality conditions yields

$$Q_k x_{k+1}^d = Q_k x_k^d - \left[\mu_k \frac{Q_k \delta_k \delta_k^T Q_k}{(\delta_k, Q_k \delta_k)} \right] (x_{k+1} - x_k^d)$$

and hence $x_{k+1}^d = x_k^d + \alpha_k \delta_k$ where

$$\alpha_k = \frac{(\delta_k, Q_k(x_{k+1} - x_k^d))}{(\delta_k, Q_k \delta_k)} \quad (13)$$

It can be shown that α_k in (13) and $\hat{\alpha}_k$ in (10c) are equivalent (see Rustem, 1990).

The inequality $(\delta_k, Q_k(x_k - x_k^d)) \leq 0$ ensures that α_k , $\mu_k \geq 0$. To demonstrate this when Q_{k+1} is given as above, we see that α_k given by (13) is equivalent to $\hat{\alpha}_k$ and this is non-negative if the above inequality and $\mu_k \geq 0$ are satisfied. When the desired value is being updated and an equivalent update to Q_k is being computed, then for $\alpha_k \geq 0$, and

$$\mu_k = - \alpha_k \frac{(\delta_k, Q_k \delta_k)}{(\delta_k, Q_k(x_{k+1} - x_k^d))} \quad (14)$$

We now show that

$$(\delta_k, Q_k(x_k - x_k^d)) \leq 0 \Rightarrow (\delta_k, Q_k(x_{k+1} - x_k^d)) \leq 0.$$

The inequality $(x_{k+1} - x_k, Q_{k+1}(x_{k+1} - x_k^d)) \leq 0$ follows from the optimality of x_{k+1} for the quadratic optimization with Q_{k+1} . Using the expression for Q_{k+1} , we have

$$\begin{aligned} 0 &\leq (x_{k+1} - x_k, Q_{k+1}(x_{k+1} - x_k^d)) \\ &= (x_{k+1} - x_k, Q_k(x_{k+1} - x_k^d)) + \frac{\mu_k}{(\delta_k, Q_k \delta_k)} \\ &\quad \times (x_{k+1} - x_k, Q_k \delta_k)(\delta_k, Q_k(x_{k+1} - x_k^d)) \end{aligned}$$

Since $(x_{k+1} - x_k, Q_k(x_k - x_k^d)) \geq 0$ follows from the optimality of x_k with Q_k , and

$$(\delta_k, Q_k(x_{k+1} - x_k)) \geq 0$$

holds if $(\delta_k, Q_k(x_k - x_k^d)) \leq 0$ (see Rustem and Velupillai, 1988b, Lemma 2) then we have $(\delta_k, Q_k(x_{k+1} - x_k^d)) \leq 0 \Rightarrow \mu_k \geq 0$ and the corresponding μ_k is given by (14). \square

The extension of the above result to nonlinear constraints is straightforward when the diagonal equivalent of a nondiagonal quadratic function is desired. However, the useful analytical equivalence of α and $\hat{\alpha}$ cannot be established exactly in the nonlinear case. The extension of the methods to nonlinear constraints can be established by invoking a mean value theorem and thereby using a local representation of the constraints (see e.g. Rustem and

*The above proposition clearly holds for $Q_k = D$ and Q_{k+1} , as given above, is D with a rank-one update and hence it is no longer, in general, diagonal.

When Q_k is singular, it can be shown that (11) can be written as $Q_k x_{k+1}^d = Q_k(x_k^d + \alpha_k \delta_k)$ from which the relevant parts of x_{k+1}^d can be recovered. For example, when $Q_k = D$, a diagonal matrix, clearly only those elements of x_{k+1}^d corresponding to nonzero diagonal elements of D can be recovered. The correspondence of $\hat{\alpha}_k$ given by (10c) and α_k can be established in the same way as in the following proof except that (10a) is used for x_{k+1} .

Velupillai, 1988b, Theorem 5). The following corollary summarizes the equivalence of the two methods when applied to nonlinear constraints.

Corollary. (The extension to nonlinear constraints). Let x_{k+1}^d and Q_{k+1} be defined by (11) and let the constraints be given by

$$G = \{x \in E^n \mid g(x) = 0\}$$

where g is twice differentiable and $g: E^n \rightarrow E^m$. Then the equivalence

$$\begin{aligned} \arg \min \{ \langle x - x_{k+1}^d, Q_k(x - x_{k+1}^d) \rangle \mid x \in G \} \\ = \arg \min \{ \langle x - x_k^d, Q_{k+1}(x - x_k^d) \rangle \mid x \in G \} \end{aligned}$$

holds for α_k given by (13).

Proof. The proof follows from the equivalence of the first order optimality conditions. \square

The extension to nonlinear constraints is thus easily implementable as the basic ingredients that enter α_k are δ_k and x_{k+1} . Both of these vectors are known when any one of the two quadratic problems have already been solved.

Proposition 2 relates the effect of a single update of Q_k that yields Q_{k+1} or a single update of x_k^d that yields x_{k+1}^d . As a corollary, we consider the sequence $\{Q_k\}$ generated by the algorithm in this section and the corresponding sequence $\{x_k^d\}$ generated by the algorithm in Section 1.

Theorem 1 (The diagonalizability of quadratic forms) Let the sequence $\{Q_k\}$ be generated by (9) and $\{x_k^d\}$ be generated by (4). Let $Q_0 = D$ then the equivalence between the diagonal and non-diagonal quadratic optimizations

$$\begin{aligned} \arg \min \{ \langle x - x_0^d, D(x - x_0^d) \rangle \mid N^1 x = b \} \\ = \arg \min \{ \langle x - x_0^d, Q_k(x - x_0^d) \rangle \mid N^1 x = b \} \end{aligned} \quad (15a)$$

holds if

$$Dx_k^d = Dx_0^d - \sum_{i=0}^{k-1} \mu_i \frac{(\delta_i, Q_i(x_k - x_0^d))}{(\delta_i, Q_i \delta_i)} Q_i \delta_i \quad (15b)$$

Proof. (15b) follows from the following equivalence of the optimality conditions of both problems

$$\begin{aligned} D(x_k - x_k^d) = Q_k(x_k - x_k^d) \\ = \left[D + \sum_{i=0}^{k-1} \mu_i \frac{Q_i \delta_i \delta_i^T Q_i}{(\delta_i, Q_i \delta_i)} \right] (x_k - x_0^d). \quad \square \end{aligned}$$

The complexity of the algorithm in Section 2 can be discussed, for Ω characterized by linear inequalities, by invoking its equivalence to the algorithm in this section and the relation of the latter to Khachian's (1979, 1980) algorithm for linear programming. As Khachian's algorithm is known to be convergent in polynomial time, its equivalence to the present algorithm would ensure the same convergence rate for the latter. It is shown in Rustem and Velupillai (1988b, Theorems 1 and 7) that the algorithm in this section is equivalent to Khachian's algorithm provided that x_0^d on the right of the equivalence (15a) is shifted, at every k , in a way that will only affect the stepsize α_k or $\hat{\alpha}_k$ above (see Rustem, 1990).

4. The arbitrariness of shadow prices

We consider the solution to Problem 2. Let there exist a quadratic objective function that reflects the policy maker's preferences and yields an optimal policy that is accepted by the policy maker. It is not necessary that this quadratic function should be specified through one of the algorithms in Sections 2 and 3. We show below that either one of these algorithms can be used, with a given quadratic function, to alter the values of the shadow prices of the problem without altering the optimal solution. The arbitrariness of the shadow prices for diagonal and nondiagonal objective functions follows from this observation. This may clearly have significant consequences in problems where shadow prices are used to indicate the "price" of scarce resources (see e.g. UNIDO, 1972). The extension of this result to general

nonlinear objective functions and nonlinear constraints is also discussed below. We begin with the non-diagonal quadratic case. Hence, we have a matrix Q_k and x^d such that

$$\arg \min \{ \langle x - x^d, Q_k(x - x^d) \rangle \mid N^1 x = b \} = x_k \in \Omega. \quad (16)$$

We do not need to assume that the inequality $\langle \delta_k, Q_k(x_k - x^d) \rangle \leq 0$ is satisfied. This is due to the transformation introduced in the following lemma.

Lemma 1. For a given δ_k , if $\langle \delta_k, Q_k(x_k - x^d) \rangle \geq 0$ and thence $\hat{\alpha}_k \leq 0$, then there exists a sufficiently small $\kappa_k \in (-\infty, 0)$ such that the transformation

$$\tilde{x}_k^d = x_k - \kappa_k(x_k - x^d) \quad (17)$$

ensures that $\arg \min \{ \langle x - \tilde{x}_k^d, Q_k(x - \tilde{x}_k^d) \rangle \mid N^1 x = b \} = x_k$ and also $\hat{\alpha}_k \geq 0$. For $\langle \delta_k, Q_k(x_k - x^d) \rangle \leq 0$, $\hat{\alpha}_k \geq 0$ with $\kappa_k = 1$ and thence $\tilde{x}_k^d = x_k^d$. Furthermore, for any sign of $\langle \delta_k, Q_k(x_k - x^d) \rangle \neq 0$, there exists a sufficiently small $\kappa_k \in (-\infty, +\infty)$ such that $\hat{\alpha}_k = 1$ and $\mu_k \geq 0$.

Proof. To show that x_k is still the solution of the revised problem, consider the solution of the latter which is denoted by \tilde{x}_k . Since $N^1[\tilde{x}_k - x_k] = 0$, we have $\tilde{x}_k - x_k = Zw$, for some vector w . Furthermore the optimality condition of the original problem yields $Q_k(x_k - x^d) - N\lambda_k = 0$. Hence, we can write the optimality condition of the transformed problem as $Q_k(\tilde{x}_k - x_k + x_k - x_k^d) - N\tilde{\lambda}_k = 0$ where $\tilde{\lambda}_k$ is the associated shadow price vector. Thus,

$$\begin{aligned} Q_k Zw = Q_k(\tilde{x}_k - x_k) = -Q_k[x_k - [x_k - \kappa_k(x_k - x^d)]] \\ + N\tilde{\lambda}_k = N[\tilde{\lambda}_k - \kappa_k \lambda_k]. \end{aligned}$$

Premultiplying by Z^T yields $Z^T Q_k Zw = 0$. As part of the sufficient conditions for optimality $Z^T Q_k Z$ is nonsingular (see e.g. Gill *et al.*, 1981), we have $w = 0$ and thence $\tilde{x}_k - x_k = Zw = 0$. To show that $\hat{\alpha}_k \geq 0$, consider

$$\hat{\alpha}_k = - \frac{\mu_k \kappa_k (\delta_k, Q_k(x_k - x^d))}{(\delta_k, Q_k \delta_k) + \mu_k (Q_k \delta_k, Z(Z^T Q_k Z)^{-1} Z^T Q_k \delta_k)}$$

Thus κ_k can be chosen to yield $\hat{\alpha}_k \geq 0$.

To show the second part of the lemma, we note that for $\langle \delta_k, Q_k(x_k - x^d) \rangle \leq 0$, the above expression yields $\hat{\alpha}_k \geq 0$ for any positive value of κ_k including $\kappa_k = 1$.

To establish the last part of the lemma, we note that $\hat{\alpha}_k = 1$ if μ_k is chosen appropriately using the above expression for $\hat{\alpha}_k$ and $\mu_k > 0$ provided κ_k is chosen to ensure a negative denominator. \square

Proposition 3 (The arbitrariness of shadow prices) Let λ_k be the vector of shadow prices of Problem (16) and assume that these shadow prices are unacceptable to the decision maker as a measure of the "price" of the corresponding scarce resources represented by the constraints. Let λ^p be the decision maker's preferred prices. Then in the algorithm of Section 3 we define δ_k as

$$Q_k \delta_k = -N[\lambda^p - \lambda_k] \quad (18)$$

then using this value, compute Q_{k+1} using (9). Given Q_{k+1} and also \tilde{x}_k^d in (17), we pass through one iteration of the algorithm to compute

$$\arg \min \{ \langle x - \tilde{x}_k^d, Q_{k+1}(x - \tilde{x}_k^d) \rangle \mid N^1 x = b \} = x_{k+1}.$$

Then, we have

$$x_{k+1} = x_k. \quad (19a)$$

$$\lambda_{k+1} = \lambda_k + \hat{\alpha}_k[\lambda^p - \lambda_k] \quad (19b)$$

where $\hat{\alpha}_k$ is given by

$$\hat{\alpha}_k = \frac{\mu_k \kappa_k (\delta_k, Q_k(x_k - x^d))}{(\delta_k, Q_k \delta_k) + \mu_k (Q_k \delta_k, Z(Z^T Q_k Z)^{-1} Z^T Q_k \delta_k)} \quad (19c)$$

* We assume that the denominator of (9) can be computed, at least approximately, for positive semi-definite Q_k . The error of such an approximation can be accounted for by an appropriate choice of μ_k .

and for μ_k and κ_k given by the last part of Lemma 1, we have

$$\lambda_{k+1} = \lambda^p. \quad (19d)$$

*Proof.** We consider the general case when Q_k is positive semi-definite. By Lemma 1, we know that x_k is unchanged by the transformation of the desired values. Since $Z^T N = 0$, expression (10a) yields

$$x_{k+1} = x_k + \alpha_k Z(Z^T Q_k Z)^{-1} Z^T [-N(\lambda^p - \lambda_k)] = x_k.$$

Similarly, expression (10b) yields

$$\begin{aligned} \lambda_{k+1} = \lambda_k - \delta_k (N^T N)^{-1} N^T (I - Q_k Z(Z^T Q_k Z)^{-1} Z^T) \\ \times [-N(\lambda^p - \lambda_k)] \end{aligned}$$

and thence we have (19b).

The last part of the proposition follows from the ability to set $\delta_k = 1$ by choosing κ_k and μ_k as established in Lemma 1. \square

Thus, it is possible to select δ_k such that it will not alter the optimal solution x_k but will change the shadow prices to achieve the preferred shadow prices, λ^p .

We consider the diagonal quadratic forms and updates of the desired values that achieve the same effect as the above proposition. We state Proposition 4 with slightly greater generality in that we do not assume that Q_k is diagonal, but simply fixed. We do not use δ_k to update Q_k and, while keeping Q_k constant, we revise the desired values as in Section 2.

Proposition 4 Let λ_k be the vector of shadow prices of problem (16) and assume that these shadow prices are unacceptable to the decision maker as a measure of the price of the corresponding scarce resources represented by the constraints. Let λ^p be the decision maker's preferred prices. Then in the algorithm of Section 2 we define δ_k as in (18) then using this value, compute

$$x_{k+1}^d = x_k^d + \alpha_k \delta_k$$

for some given value of α_k . We pass through one iteration of the algorithm to compute

$$\arg \min \{ (x - x_{k+1}^d, Q_k(x - x_{k+1}^d)) \mid N^T x = b \} = x_{k+1}$$

Then, we have $x_{k+1} = x_k$ and

$$\lambda_{k+1} = \lambda_k + \alpha_k [\lambda^p - \lambda_k], \quad (20)$$

and with the choice $\alpha_k = 1$, we have $\lambda_{k+1} = \lambda^p$.

Proof. Let x^d be the desired value used in (16). Using the same arguments as in (7), we have

$$x_{k+1} = x_k + ZW = -Z(Z^T Q_k Z)^{-1} Z^T Q_k [x_k - x_k^d - \alpha_k \delta_k]$$

The optimality condition of (16) yields

$$Q_k(x_k - x^d) - N\lambda_k = 0 \quad (21)$$

and δ_k is given by (18) and as $Z^T N = 0$, we have $x_{k+1} - x_k = 0$.

To establish (20) consider (8) which can be rewritten as

$$\lambda_{k+1} = (N^T N)^{-1} N^T Q_k [x_{k+1} - x_k + x_k - x_k^d - \alpha_k \delta_k].$$

Using (21), (18) and $x_{k+1} - x_k = 0$, this yields (20).

The choice of $\alpha_k = 1$ immediately yields $\lambda_{k+1} = \lambda^p$ using the same arguments as above. \square

The extension of the above arbitrary nature of the shadow prices can also be extended to nonlinear constraints and this is considered in Rustem (1990). In Theorem 2, we discuss the case when the desired values are updated to alter the shadow prices of the nonlinear problem.

* In this result Q_k is not assumed to be positive definite. All that is needed is that $(\delta_k, Q_k(x_k - x^d)) \neq 0$. Provided this condition is satisfied, we can arbitrarily change the shadow prices.

Theorem 2. Arbitrariness of shadow prices for nonlinear problems when the desired values are updated. Let

$$Q_k \delta_k = -\nabla g(x_k) [\lambda^p - \lambda_k]$$

then

$$x_{k+1}^d = x_k^d + \delta_k$$

ensures that

$$\arg \min \{ (x - x_k^d, Q_k(x - x_k^d)) \mid x \in G \} = x_k$$

and

$$\arg \min \{ (x - x_{k+1}^d, Q_k(x - x_{k+1}^d)) \mid x \in G \} = x_{k+1} = x_k$$

and furthermore, $\lambda_{k+1} = \lambda^p$.

Proof. From the optimality conditions of both problems we have

$$\begin{aligned} Q_k(x_k - x_k^d + \delta_k) + Q_k(x_k - x_k^d) + \nabla g(x_k) [\lambda^p - \lambda_k] \\ = -\nabla g(x_k) \lambda^p \end{aligned}$$

Hence the solution of the second problem is given by x_k and λ^p . \square

Conclusions

The possibility of constructing quadratic objective functions with diagonal weighting matrices, or Hessians, is desirable both in terms of computational convenience and interpretability. The extension to nonlinear constraints permits the wider applicability of the results.

The arbitrariness of shadow prices seems to be an intuitively obvious concept. Yet there has been numerous attempts in the past to use them as the price of scarce resources or activities. The two procedures discussed, one by the updating of the weighting matrix and the other by the updating of the desired values, actually construct arbitrary shadow price values without altering the optimal solution.

Acknowledgements. The financial support of ESRC is gratefully acknowledged. The paper has also benefited from the comments of two anonymous referees and from valuable discussions with Robin Becker.

References

- Dreze, J. P. (1982). On the choice of shadow prices for project evaluation. Indian Statistical Institute Discussion Paper No. 16.
- Engwerda, J. C. and P. W. Otter (1989). On the choice of weighting matrices in the minimum variance controller. *Automatica*, **25**, 279-285.
- Fiacco, A. N. and G. P. McCormick (1968). *Nonlinear Programming Sequential Unconstrained Minimization Techniques*. Wiley, New York.
- Geoffrion, A. M. (1971). Duality in nonlinear programming: A simplified applications oriented development. *SIAM Review*, **13**, 1-37.
- Gill, P. E., W. Murray and M. H. Wright (1981). *Practical Optimization*. Academic Press, London.
- Hughes-Hallett, A. J. and H. J. B. Rees (1983). *Quantitative Economic Policies and Interactive Planning*. Cambridge University Press, Cambridge, U.K.
- Khachian, L. G. (1979). A polynomial algorithm in linear programming. *Soviet Math Doklady*, **20**, 191-194.
- Khachian, L. G. (1980). Polynomial algorithm in linear programming. *USSR Comput. Math. Math. Phys.*, **20**, 53-72.
- Polak, F. (1971). *Computational Methods in Optimization*. Academic Press, New York.
- Ponstein, J. (1983). Applying some modern developments to choosing your own Lagrange multipliers. *SIAM Review*, **25**, 183-199.
- Rustem, B. (1990). On the diagonalizability of quadratic forms and the arbitrariness of shadow prices. In N. Christodoulakis (Ed.), *Modelling and Control of National Economies*. Pergamon, Oxford.
- Rustem, B. and K. Velupillai (1988a). Constructing

<p>objective functions for macroeconomic decision models: A formalization of Ragnar Frisch's approach. PROPE Discussion Paper No. 69. Imperial College, London.</p> <p>Rustem, B. and K. Velupillai (1988b). Constructing objective functions for macroeconomic decision models: On the complexity of the policy design process. PROPE Discussion Paper No. 84. Imperial College, London.</p>	<p>Rustem, B. K., K. Velupillai and J. Westcott (1978). Respecifying the weighting matrix of a quadratic objective function. <i>Automatica</i>, 14, 567-582.</p> <p>UNIDO (1972). <i>Guidelines for Project Evaluation</i>. United Nations Publications, UNIDO, Vienna.</p>
---	---

Brief Paper

A Hierarchical-multiobjective Framework for Risk Management*

YACOV Y. HAIMES†‡ and DUAN LI†

Key Words—Large-scale systems; risk; multiobjective optimization; hierarchical decision making; reliability.

Abstract—The management of risk is addressed in this paper within the broad hierarchical-multiobjective framework. Such a framework incorporates the hierarchical nature of the decision-making process; the multiple decision-makers at the various levels of the system's hierarchy; the multiobjective nature of large-scale systems; and the quantitative/empirical and the qualitative/normative/judgmental aspects. Three major topics dominate the methodological components of the paper: hierarchical-multiobjective coordination; risk of extreme events and impact analysis. Various application problems are used as a vehicle to communicate the methodological framework with the readers.

1. Introduction

RISK, A MEASURE of the probability and severity of adverse effects—has become during the last decade a subject of intense study by scholars from diverse disciplines. In particular, the process of risk assessment and management, which encompasses several steps that include the identification, measurement, quantification, evaluation and management/control of risk, has gained cross-disciplinary attention that spans engineering; the natural, behavioral and social sciences; law; medicine; and business administration. Previous research in stochastic modeling and optimization, nevertheless, by and large addressed the quantitative, empirical aspects of risk management separately from the qualitative, normative and judgmental considerations, which are the driving forces that ultimately influence the decision-making process. The focus on the former and the de-emphasis of the latter have been counter to the holistic foundations upon which systems engineering is grounded. This paper attempts to capture some specific characteristics of risk assessment and management within the overall decision-making process. If one accepts the premise that the decision-making process itself is driven by multiple conflicting and noncommensurate objectives and that large-scale organizational and technological systems are characterized by inherent hierarchical structures and a hierarchy of decision-makers, then the process of risk assessment and management can be best understood, and thus modeled, via a hierarchical-multiobjective framework. Furthermore, the thesis of this paper is also grounded on the

premise that risk management should be an integral part of technology management, not a vacuous afterthought. Three major topics dominate the methodological components of the paper: hierarchical-multiobjective coordination; risk of extreme events and impact analysis. Applications from the areas of the maintenance of infrastructure and flood warning and evacuation systems are used as a vehicle to communicate the methodological framework to the readers.

2. Multiple objective aspects

Lowrance (1976) makes a clear distinction between risk and safety. Measuring risk is an empirical, scientific activity (e.g. measuring the probability and severity of harm), on the other hand, judging safety, which is a normative, political activity, is judging the acceptability of risks. The distinction between risk and safety is at the heart of multiobjective trade-off analysis. Clearly, the determination of how safe is safe enough requires a balance among all costs, benefits and risks—attributes that are measured and perceived in different, noncommensurate units. For example, a classical multiobjective analysis might involve trading off the spending of an additional \$1000 per day for the reduction of one part per billion (ppb) of trichloroethylene (TCE) in a municipality's contaminated groundwater (given that the contamination level is already at 10 ppb and \$2 million has already been spent in cleaning up the aquifer that supplies drinking water to a community of 20,000 people). Such an analysis is also a risk management problem where the empirical act of measuring the risk of contamination must be followed by the normative step of determining the level of risk that is deemed acceptable by the decision-maker(s).

During the last two decades, the field of multiple criteria decision-making (MCDM) has grown by leaps and bounds, developing from a primarily utility-theory-oriented school to a diverse and balanced philosophical interpretation of utilities, attributes and objectives. More specifically, the nonutilitarian school of thought advances the premises not only that it is extremely difficult if not actually impossible in practice to model the preferences of a decision-maker (or of a group of decision-makers) through a utility function, but it is also not needful or desirable to do so. This premise is supported by the observation that the utility of a decision-maker is likely to be highly nonlinear and dynamic—constantly influenced by transient and exogenous elements that cannot be accounted for quantitatively.

The nonutilitarian multiobjective trade-off approach [e.g. the use of the surrogate worth trade-off (SWT) method (Haimes and Hall, 1974)] is particularly appropriate for risk management because the acceptability of risk is invariably driven by a host of perceptions, heuristics, and biases—a reality that tends to influence and shift the judgement of an acceptable risk level from an absolute venue toward a relative one that is subject to continuous changes and modifications (Kahmaman *et al.*, 1982). This fact coupled with the hierarchical structure of most large-scale systems and the hierarchical nature of the decision-making process (as will be discussed subsequently) renders the hierarchical-multiobjective framework for risk management a natural imperative.

* Received 24 December 1989; revised 29 July 1990; received in final form 5 September 1990. The original version of this paper was presented in a plenary session at the IFAC/IFORS/IMACS Symposium on Large Scale Systems: Theory and Applications, 29–31 August 1989, Berlin, Germany. The Published Proceedings of this IFAC Meeting may be ordered from: Pergamon Press plc, Headington Hill Hall, Oxford OX3 0BW, U.K. This paper was recommended for publication in revised form by Editor A. P. Sage.

† Center for Risk Management of Engineering Systems and the Department of Systems Engineering, University of Virginia, Charlottesville, Virginia 22901, U.S.A.

‡ Author to whom all correspondence should be addressed.

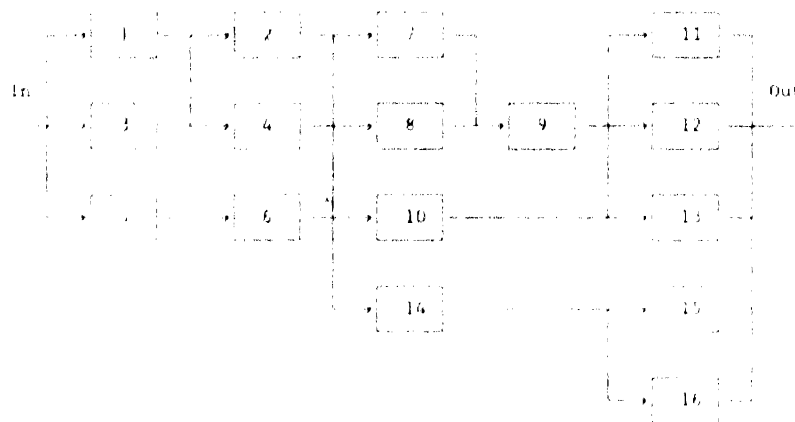


FIG. 1. Water distribution system.

3. Hierarchical aspects

Most organizational as well as technological systems are hierarchical in nature, and thus the management of risk of such systems is necessarily driven by this hierarchical reality and must be responsive to it. The risks associated with each subsystem within the hierarchical structure contribute to and ultimately determine the risks of the overall system. The distribution of risks within the subsystems often plays a dominant role in the allocation of resources within the organizational or technological system. This is manifested in the quest to achieve a level of risk that is deemed acceptable in the normative-judgmental decision-making process, when the trade-offs among all the costs, benefits and risks are considered.

Perhaps one of the most valuable and critical contributions of the hierarchical-multiobjective framework to risk assessment and management is its ability to facilitate the evaluation of the risks associated with each subsystem and their corresponding contribution to the overall risks of the total system. In the planning, design or operational mode, the ability to model and quantify the risks contributed by each subsystem to the overall system markedly facilitates the identification, quantification and evaluation of risk. In particular, the ability to model the intricate relationships among the various subsystems and to account for all relevant and important elements of risk and uncertainty renders the modeling process more tractable and the risk assessment process more representative and encompassing. Consider, for example, the problem of maximizing the availability measure of a maintainable infrastructure system. It is known that a given level of availability measure can be achieved by many different combinations of reliability and maintainability. Reliability is defined here as the probability that the system is operational in a given time period. The system's reliability can be improved by applying a certain class of preventive maintenance policies, such as an "age" policy or a "block" policy. Maintainability is defined here as the probability of the duration of the system's downtime resulting from either scheduled or emergency shutdowns. The system's reliability or the maintainability of each of its subsystems can be independently improved if there is no budget constraint. In most real-world situations, however, a resource limitation usually acts as the driving force, and trade-offs thus exist between the reliability and the maintainability of the overall system.

Hierarchical control, when applied to risk management systems, induces a harmonizing effect over the subsystems and contributes to the holistic approach within which the overall system is viewed. Fault tree analysis (U.S. Nuclear Regulatory Commission, 1981), for example, is a widely used analytical tool that decomposes the overall reliability problem into several levels of reliability problems and systematically calculates the failure rate of the overall (top) event from the lower level to the upper level. Studies aiming at developing risk management strategies using decomposition and higher-level coordination are currently under way. Dealing with a low-dimensional multiobjective optimization

problem and identifying the impact of the subsystems' reliability on the overall system's performance, a preferred Pareto optimal solution of a large-scale overall system can be reached by introducing coordination among the subsystems.

4. Hierarchical-multiobjective framework for large-scale infrastructure problems

In the preface to a major study by the U.S. National Academy of Engineering (1988) on infrastructure, Robert R. White states that "infrastructure is the term applied to large-scale engineering systems and includes an array of public works, such as roads, bridges, and sewer systems, as well as privately managed utilities such as electric power and telephone service." Thus, the fundamental characteristics of infrastructure problems lie in their large number of components and subsystems. Most water distribution systems, for example, must be addressed within a framework of large-scale systems. In addition, a hierarchy of institutional and organizational decision-making structures (e.g. federal, state, county and city) is often involved in determining the best replacement/repair strategy. A certain degree of coupling exists among the subsystems (e.g. the overall budget constraint imposed on the overall system), and this further complicates the management of such systems. Different replacement/repair strategies for different subsystems often have different impacts on the overall water distribution system, the needs for the resources and their appropriate allocation have a diverse impact on the reliability of the overall system.

The modeling of deteriorating water distribution systems is a focal issue in large-scale infrastructure problems (Andreou *et al.*, 1987; Mays and Cullinane, 1986). A water distribution system may consist of many subsystems. The type of complex water distribution system considered in this paper is a series-parallel network. One example is given in Fig. 1.

The unreliability of an overall water distribution system can be expressed as a function of the unreliabilities of the water distribution subsystems

$$F_i = F_i(f_1, f_2, \dots, f_N) \quad (4.1)$$

where F_i is the unreliability of the overall water distribution system and f_i , $i = 1, 2, \dots, N$, is the unreliability of the i th water distribution subsystem. For the example given in Fig. 1, one possible decomposition may be as follows: Components 1, 2, 3 and 4 constitute subsystem 1; components 5 and 6, subsystem 2; components 7, 8, 9 and 10, subsystem 3; components 11, 12 and 13, subsystem 4; and components 14, 15 and 16, subsystem 5. The associated fault tree is given in Fig. 2. The overall system's unreliability is given by

$$F_i = 1 - [1 - f_1 f_2] \{1 - [1 - (1 - f_3)(1 - f_4)] f_5\}.$$

In most cases, the optimization of a water distribution system is difficult to handle as a whole. Hierarchical (multilevel)-multiobjective approaches (Tarvainen and Haimes, 1982; Li and Haimes, 1987a, 1988; Haimes and Li,

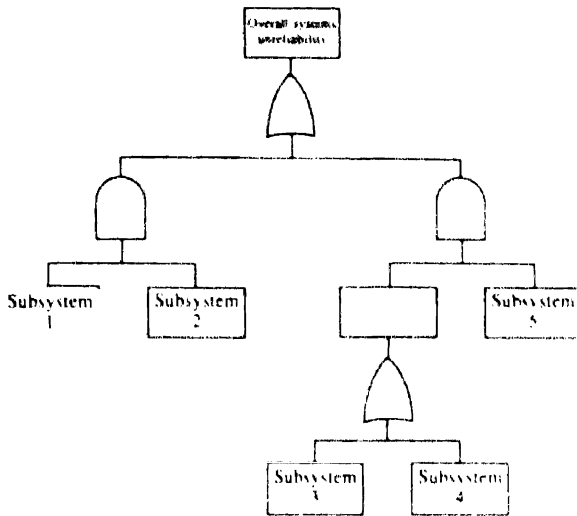


FIG. 2. Fault tree of the water distribution system

1988; Haines *et al.*, 1990b) solve large-scale multiobjective optimization problems by decomposition and upper-level coordination. In general, the structural nature of multilevel decomposition shows the following advantages.

- (1) Decomposition methods can reflect the internal hierarchical nature of large-scale multiobjective systems.
- (2) Trade-off analyses can be performed among subsystems and the overall system, and
- (3) Through decomposition, the complexity of a large-scale multiobjective system can be relaxed by solving several smaller subproblems.

Assume that the following water distribution system consists of N subsystems. The overall system's multiobjective optimization problem is posed as follows:

$$\min [F_1(C_1, \dots, C_N), F_2(f_1, \dots, f_N)]^T \quad (4.2a)$$

$$\text{subject to } y_i = H_i(x_i, m_i) \quad i = 1, 2, \dots, N \quad (4.2b)$$

$$g_i(x_i, m_i, y_i) \leq 0 \quad i = 1, 2, \dots, N \quad (4.2c)$$

$$\sum_{i=1}^N B_{ij} y_j \quad i = 1, 2, \dots, N \quad (4.2d)$$

where $C_i = C_i(x_i, m_i, y_i)$ and $f_i = f_i(x_i, m_i, y_i)$ are the cost function and the unreliability of subsystem i , respectively;

$F_1 = \sum_{i=1}^N C_i$ is the overall cost and $F_2 = F_2(f_1, \dots, f_N)$ is the

overall system's unreliability; y_i is the output of subsystem i , m_i is the control of subsystem i ; x_i is the interaction input of subsystem i ; H_i represents the system's equation of subsystem i ; g_i represents the constraints of subsystem i ; and B_{ij} is a matrix representing the interaction between subsystem i and the other subsystems.

The optimization problem for a water distribution system will be used as a vehicle to present the hierarchical-multiobjective framework for risk management. At the lower level, the overall optimization problem can be decomposed into several smaller water distribution subsystems, which are interconnected by fixed values of coordination variables set by the upper level. The coordination variables often correspond to economic meaning, e.g. shadow prices. At the upper level, the coordinator adjusts the coordination variables according to the trade-off values among the subsystems and the overall systems.

Assume that the upper-level decision-maker's preferences expressed in terms of trade-offs are available. The internal indifference trade-off vector λ^i in subsystem i is then obtained by mapping the indifference trade-off vector, $[1, \lambda_{12}(F_1, F_2)]$, of the overall system into each subsystem:

$$\lambda^i = \left[1, \lambda_{12}(F_1, F_2) \frac{\partial F_2}{\partial f_i} \right]^T \quad (4.3)$$

Specific values y_i^k are selected for output vectors y_i at the upper level at iteration k in the feasible decomposition

method. Then the overall system can be decomposed into N subsystems,

$$\min [C_i(x_i, m_i, y_i^k), f_i(x_i, m_i, y_i^k)]^T \quad (4.4a)$$

$$\text{subject to } y_i^k = H_i(x_i, m_i) \quad (4.4b)$$

$$x_i = \sum_{j=1}^N B_{ij} y_j^k \quad (4.4c)$$

$$g_i(x_i, m_i, y_i^k) \leq 0 \quad (4.4d)$$

where the trade-off vector λ^i generated by equation (4.3) is imposed on the subsystems by the higher-level coordinator to generate the preferred solution of the multiobjective optimization problem given in (4.4).

In the nonfeasible method, the Lagrangian multipliers v_i associated with (4.2d) are used as the coordination parameter. Assume that superscript k is used to indicate values assigned by the upper level at iteration k . The optimization problem for each subsystem i then becomes

$$C_i(x_i, m_i, y_i) + \sum_{j=1}^N (v_j^k)^T B_{ij} \left[y_j - (v_j^k)^T x_i \right] \quad (4.5a)$$

$$f_i(x_i, m_i, y_i) \quad (4.5b)$$

$$\text{subject to } y_i = H_i(x_i, m_i) \quad (4.5c)$$

$$g_i(x_i, m_i, y_i) \leq 0 \quad (4.5d)$$

where the trade-off vector λ^i generated by (4.3) is imposed on the subsystems by the higher-level coordinator to generate the preferred solution of the multiobjective optimization problem given in (4.5).

Based on the solutions generated at the lower level, the upper level modifies the values of the coordination variables and sends them back to the lower level. This iteration process is repeated until all optimal conditions are met.

5. Risk of extreme events

5.1. Overview. By its nature, an uncertain event defies our ability to characterize its state in terms of one single definite number. Indeed, the modeling of a random variable through a probability density function (pdf) constitutes the best quantitative characterization of an uncertain event. In effect, the pdf assigns a probability for the occurrence of the uncertain event at a given level within a prespecified domain. In our quest to simplify the decision-making process under uncertainty, it has been a common practice to use the expected value of the random variable as the sole criterion that represents the probability and severity of the random variable. Often, the expected value is used as the risk measure associated with the random variable.

Reconsider, for example, the contamination of a groundwater system with the known carcinogen trichloroethylene (TCE). Since the concentration of TCE in the groundwater cannot be known for certainty, a pdf can be generated on the basis of a sampling and modeling effort. Let X be the random variable representing the concentration of TCE in parts per billion (ppb); let $p_i(x)$ be the probability density function (pdf) of the TCE concentration; and let $P_i(x)$ be the cumulative distribution function (cdf) of the TCE concentration. Then, the expected value $E(X)$ of the concentration of TCE in the groundwater system is given by:

$$E(X) = \int_0^\infty xp_i(x) dx \quad (5.1)$$

Note that the expected value is a mathematical artifice that commensurates concentrations (events) having high values and low probabilities with concentrations having low values and high probabilities. Such an averaging process, while helpful in some respects, can distort the true danger (risk) of extreme values (of TCE concentration), leading to complacency and ultimately to disasters. Experience has shown, time and again, that when the expected value is used as the sole index for risk, it often leads to the do-nothing option. On the other hand, when a different index, the conditional expected value (an index that measures the probability and severity given that the event occurs in a

specified range of probability or a range of severity, e.g. TCE concentration) is used, the do-nothing option often becomes an inferior one. The expected value of adverse effects, which has been the most commonly used measure of risk, is in many cases inadequate, since this scalar representation of risk commensurates events that correspond to all levels of losses and their associated probabilities. The common expected-value approach is particularly deficient for addressing extreme events, since these events are concealed during the amalgamation of events of low probability and high consequence and events of high probability and low consequence. The partitioned multiobjective risk method (PMRM) and its extensions (using results from the statistics of extremes) provide a valuable tool in the quantification and evaluation of risk focusing on extreme and catastrophic events (Asbeck and Haimes, 1984; Karlsson and Haimes, 1988a, 1988b; Mitsiopoulos and Haimes, 1989).

The PMRM is a risk analysis method developed for solving multiobjective problems with a probabilistic nature. Instead of using the traditional expected value, the PMRM generates a number of conditional expected risk functions, given that the damage falls within specific probability ranges (or damage ranges). Assume the damage (TCE concentration in the groundwater case) can be represented by a continuous random variable X with a known probability density function $p_x(x; y_j)$, where $y_j, j = 1, \dots, q$, is a control policy. The PMRM partitions the probability axis into three ranges. Denote the partitioned points on the probability axis by $\alpha_i, i = 1, 2$. For each α_i and each policy y_j , it is assumed that there exists a unique damage β_{ij} such that

$$P_x(\beta_{ij}; y_j) = \alpha_i \quad (5.2)$$

where P_x is the cumulative distribution function of X . These β_{ij} (with β_{0j} and β_{1j} representing, respectively, the lower bound and upper bound of the damage) define the conditional expectation as follows:

$$f_i(x_i) = \begin{cases} \int_{\beta_{0j}}^{\beta_{1j}} xp_x(x; y_j) dx & i = 2, 3, 4 \\ \int_{\beta_{1j}}^{\beta_{2j}} xp_x(x; y_j) dx & j = 1, \dots, q \end{cases} \quad (5.3)$$

where f_2, f_3 and f_4 represent the risk with high probability of exceedance and low damage, the risk with medium probability of exceedance and medium damage, and the risk with low probability of exceedance and high damage, respectively.

The unconditional expected value of X is denoted by $f_x(y_j)$. The relationship between the conditional expected values (f_2, f_3, f_4) and the unconditional expected value (f_x) is given by

$$f_x(y_j) = \theta_2 f_2(y_j) + \theta_3 f_3(y_j) + \theta_4 f_4(y_j) \quad (5.4)$$

where $\theta_i, i = 2, 3, 4$, is the denominator of (5.3).

Combining one of the generated conditional expected risk functions or the unconditional expected risk function with the cost function, f_1 , creates a set of multiobjective optimization problems:

$$\min [f_1, f_i] \quad i = 2, 3, 4, 5 \quad (5.5)$$

subject to the system's constraints.

Solving the family of the above multiobjective optimization problems offers more information about the probabilistic behavior of the problem than the single multiobjective formulation of minimizing the cost and the unconditional expected risk function, $\min [f_1, f_x]$. The trade-offs between the cost function f_1 and any risk function $f_i, i \in \{2, 3, 4, 5\}$ enable decision-makers to evaluate the marginal cost of a small reduction in the risk objective given a particular level of risk assurance. The relationship of the trade-offs between the cost function and the various risk functions is given by

$$1/\lambda_{1i} = \theta_2/\lambda_{12} + \theta_3/\lambda_{13} + \theta_4/\lambda_{14} \quad (5.6)$$

where

$$\lambda_{1i} = -\partial f_1 / \partial f_i, \quad \lambda_{1i} > 0, \quad i = 2, 3, 4, 5 \quad (5.7)$$

is the trade-off value between f_1 and f_i in (5.5). Knowledge of

this relationship among the marginal costs is useful for the decision-makers to determine an acceptable level of risk. Any multiobjective optimization method, e.g. the surrogate worth trade-off (SWT) method (Haimes and Hall, 1974; Chankong and Haimes, 1983), can be applied at this stage.

While the multiobjective nature of risk-based decision-making is obvious from the preceding discussion, its hierarchical nature deserves more elaboration (Li and Haimes, 1987c). Indeed, rarely are policy options on important and encompassing issues formulated, traded off, evaluated, and finally decided upon at one single level in the hierarchical decision-making process. Rather, a hierarchy that represents various constituencies, stakeholders, power brokers, advisors and administrators and a host of shakers and movers constitutes the true players in the complex decision-making process. Relating to the groundwater contamination problem in the U.S. context, one may view the U.S. Environmental Protection Agency (EPA) as representing the higher level in the hierarchical-multiobjective decision-making framework (notwithstanding the fact that the USEPA itself has its own hierarchical decision-making structure). The lower levels include the Department of Natural Resources of each of the 50 states (or the state's EPA) down to the regional, state and local levels. Furthermore, concerns about the expected value of contamination and the conditional expected value of extreme contamination with TCE vary within each level of the hierarchy; consequently, the risk measures vary correspondingly. At the upper level, the USEPA may consider long-term effects and a much broader geographical area, whereas at a lower level, a local government may consider a shorter time horizon and a much more localized geographical region.

5.2. Classification of risk-control systems. The classification of systems on the basis of their response to extreme events can be valuable to the decision-makers in their qualitative/normative evaluation of the acceptability of risk (Li and Haimes, 1990). Such a classification might also provide an insight into the impacts that current decisions might have on future risk management options.

If a control has the same impact on the risk expectation and on the conditional risk expectation of the extreme events, we call this class of system "systems with neutral risk control". In other words, the control is "neutral" in terms of minimizing the expected value and in minimizing the expected extremes.

If a control has different impacts on the risk expectation and on the conditional risk expectation of extreme events, we call this class of system "systems with risk-manipulatable control". In most cases, risk management strategies make a greater reduction in the measure of the conditional expected extreme than in the measure of the expected value. Furthermore, for systems with risk-manipulatable control, if the minimization of the expected value is consistent with the minimization of the associated variance, we call this type of system "systems with mean-variance consistency risk control".

Finally, when there exists an inconsistency between the mean and the variance, minimizing the expected value and the expected extreme may conflict with each other. We call this class of system "systems with mean-variance inconsistency risk control".

The classification of risk-control systems can also prove to be useful for researchers in the field; it provides a taxonomy that is based not only on the nature of risk but also on its impact.

6. Impact analysis

6.1. Overview. Good technology management necessarily incorporates good risk management practices. Determining the impacts of current decisions on future options, however, constitutes what might be termed as the imperative in good decision-making. Managers, public officials and other decision-makers are commonly rewarded, promoted and otherwise honored not because of the large number of optimal decisions that they make during their tenure in

office; rather, they are acknowledged primarily and dominantly for the few disastrous decisions that they make. This trend explains, to a large extent, the conservative and often rigid attitude of bureaucrats who avoid untested paths and experimental options. The ability to model and assess the impacts of current risk-based decisions on the state of the system in the future can thus prove to be a potent tool in decision-making. This ability is particularly valuable for risk-based decision-making associated with dynamic systems, where decision-makers often need to balance short-term with long-term objectives.

In this sense, impact analysis is paramount to "looking before you leap". In particular, stage trade-offs in dynamic systems are needed to measure the impact of the variations of the objective functions at the present stage upon the levels of the objective functions at the remaining stages. Impact analysis thus provides useful information that might avoid adverse and irreversible consequences resulting from what might be perceived as an optimum present decision.

Consider, for example, the risk management problem associated with flood warning and evacuation systems (Haimes *et al.*, 1990a). In general, such systems can be decomposed into a two-level structure. Two subsystems constitute the lower level—the forecast subsystem and the community response subsystem. Based on hydrological and meteorological information and observations, the forecast subsystem calculates the forecasted flood crest. By its nature, there can be no perfect forecast system, and two types of errors of forecasts occur. Type I errors are the missed forecasts and Type II are the false alarms. The performance of a forecast system can be best judged by some statistical measures. The fraction of people in the community who respond to the flood warning constitutes the state variable of the community response system and is dependent on the past performance of the forecast system. Type II errors (false alarms) have a "cry wolf" effect and markedly reduce the system's credibility (thus decreasing the future response of the population to warnings or to evacuations). The general interaction between the forecast subsystem and the response subsystem is given in Fig. 3. The task of the second-level coordination is to set a warning threshold, which can be preassigned such that a flood warning will be issued when the forecasted flood crest is higher than the warning threshold. If the warning threshold is set higher, there will be a few false alarms and more missed forecasts, and vice versa. Type I and II errors have different impacts on flood-loss reduction. A missed forecast results in an immediate flood loss, while a false alarm reduces the population response fraction in the future. A lower response fraction will cause a higher flood loss; thus, there exist trade-offs between short-term and long-term risk management objectives.

6.2. Hierarchical impact analysis for flood warning systems. The flood warning and evacuation system will be used as a vehicle to demonstrate the hierarchical-methodological framework for impact analysis. Define H to be a random variable which represents the actual flood crest and S to be a random variable which represents the forecasted flood crest. If the prior probability density function of the flood crest is denoted by $g(h)$ and the conditional probability density function of s , given h , is denoted by $f(s|h)$, then the posterior probability density function of h , given forecast s , is

$$f(h|s) = f(s|h)g(h)/k(s) \quad (6.1)$$

where $k(s)$ is the marginal probability density function of

forecast s ,

$$k(s) = \int_0^\infty f(s|h)g(h)dh \quad (6.2)$$

Based on hydrological and climatic information, both observed and historical, the forecast subsystem provides the forecasted value s of the flood crest. From the Bayesian formula given in (6.1), the prior representation of the uncertainty associated with the flood crest is then updated to a posterior form $f(h|s)$.

The flood warning threshold s^* is introduced in the following context:

- (a) No flood warning will be issued if $s < s^*$, and
- (b) A flood warning will be issued if $s \geq s^*$

In other words, a flood warning will be issued only when the forecasted flood crest s exceeds a preassigned threshold level s^* . The selection of the flood warning threshold directly affects the evacuation action which is implemented by the community response subsystem.

Assume that the elevation of the floodplain zone under consideration is y ; the probability then that this zone will be flooded, conditioned on the forecast s , is $P(h \geq y|s)$. There exist four possible outcomes that follow a flood warning decision: a correct warning that is a warning followed by a flood, a false warning that is a warning not followed by a flood, a missed warning that is a flood event not preceded by a warning, and a correct quiet that is an event of no warning and no flood. The probability of a correct warning is

$$P_{11}(s^*, y) = \int_{s^*}^\infty P(h \geq y|s)k(s)ds \quad (6.3)$$

The probability of a false warning is

$$P_{10}(s^*, y) = \int_0^{s^*} P(h \geq y|s)k(s)ds \quad (6.4)$$

The probability of a missed flood warning is

$$P_{01}(s^*, y) = \int_0^{s^*} P(h < y|s)k(s)ds \quad (6.5)$$

The probability of acting correctly in not issuing a flood warning is

$$P_{00}(s^*, y) = \int_{s^*}^\infty P(h < y|s)k(s)ds \quad (6.6)$$

It is clear from (6.3)–(6.6) that the value of the selected threshold s^* plays a key role in determining the probabilities of Type I and Type II errors. If the threshold s^* is set lower, the forecast will have a lower probability value, P_{01} , of a Type I error and a higher probability value, P_{10} , of a Type II error. If the threshold s^* is set higher, the forecast will have a higher probability value, P_{01} , of a Type I error and a lower probability value, P_{10} , of a Type II error.

Denote the fraction of people in the community who respond to a call of evacuation by α_T for the T th flood event. If a past flood event has been predicted, then the confidence in the forecast system will increase, and thus, future rates of response will also increase. On the other hand, a Type II error will decrease confidence in the forecast system, thereby decreasing future rates of response. The experience of a missed warning will decrease people's confidence in a flood warning system and increase people's alertness to the possibilities of future floods. For simplicity, it is reasonable to assume here that the response fraction will remain unchanged after a missed warning has been experienced. It is also reasonable to assume that a correct quiet does not change the response fraction in the future. In view of the above discussion, the fraction α_T can thus be assumed to evolve dynamically as a controlled stochastic process:

$$\alpha_{T+1} = \begin{cases} \alpha_T + c_1(1 - \alpha_T) & \text{with prob. } P_{11}(s_T^*, y) \\ \alpha_T & \text{with prob. } P_{00}(s_T^*, y) + P_{01}(s_T^*, y) \\ c_2\alpha_T & \text{with prob. } P_{10}(s_T^*, y) \end{cases} \quad (6.7)$$

where the values of c_1 and c_2 are between zero and one and

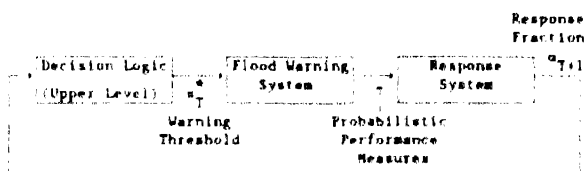


FIG. 3. Interaction between forecast and response subsystems.

can be determined using identification methods based on historical data.

The flood warning threshold cannot be selected in isolation at each stage since the decision-maker must balance the desire for high present flood-loss reduction with the possibility of high future flood loss. A multiobjective multistage optimization model can be adopted to find the optimal values for the flood warning threshold at various stages. Evaluating the trade-off between short- and long-term effects leads to an acceptable balance between the expected loss reduction at the current stage and the future response fraction that is the key element in flood-loss reduction.

Assume that there are N successive flood events in the time horizon under consideration. Denote the expected property-loss reduction at the T th flood event by $f_1^T(\alpha_T, s_T^*, y)$, and the expected life-loss reduction at the T th flood event by $f_2^T(\alpha_T, s_T^*, y)$. The maximization of three noncommensurate/conflicting objective functions is considered. The first objective is to maximize the sum of the expected property-loss reductions of all flood events over the time horizon under consideration; the second objective is to maximize the sum of the expected life-loss reductions of all flood events over the time horizon under consideration; and the third objective is to maximize the forecast system's credibility, which is implicitly expressed by $E(\alpha_{N+1})$, the expected fraction of people who respond to the warning beyond the time horizon under consideration. Mathematically, this overall multiobjective optimization problem can be posed as follows:

$$\max_{\alpha} \begin{bmatrix} f_1 = E \left\{ \sum_{T=1}^N f_1^T \right\} \\ f_2 = E \left\{ \sum_{T=1}^N f_2^T \right\} \\ f_3 = f_3^N = E(\alpha_{N+1}) \end{bmatrix} \quad (6.8)$$

subject to (6.7).

The multistage multiobjective optimization problem in (6.8) can be effectively solved by multiobjective dynamic programming approaches, such as the envelope approach by Li and Haimes (1987b, 1988). The set of noninferior solutions provides the decision-maker the best solutions to balance the short- and long-term objectives and to avoid adverse consequences resulting from what might be perceived as an optimum present decision.

7. Conclusions

Risk and uncertainty are fundamental elements of modern life; they are ever-present in the actions of human beings, and are frequently magnified in large-scale technological systems. Engineering systems, for example, are almost always designed and operated under conditions of risk and uncertainty and are often expected to achieve multiple and conflicting objectives. Success is gauged by engineering, economic, legal and social criteria. To make a rational choice, the decision-maker must evaluate the alternatives in light of all these criteria, analyze trade-offs between them and make a final decision that combines engineering analysis with societal preferences. These complex and interrelated forces necessitate that risk and uncertainty be managed effectively within a holistic framework. The proposed hierarchical-multiobjective framework constitutes the building blocks for risk management in such a holistic framework. Those private and public organizations that can successfully address the risk inherent in their business—whether future product design, resource availability, natural forces, market changes or the reliability of man/machine systems—will dominate the technological market.

Acknowledgements—Financial support for the research was provided, in part, by the National Science Foundation, Grant No. CES-8617984, under the project title "Hierarchical-multiobjective management of large scale infrastructure"; the National Aeronautics and Space Administration, Contract No. NASA-4311, under the project title "Integration of the partitioned multiobjective risk method (PMRM)

and fault-tree analysis"; and the Institute for Water Resources, U.S. Army Corps of Engineers. The editorial work of Virginia Benade and Gail Hyder Wiley is very much appreciated.

References

- Andreou, S. A., D. H. Marks and R. M. Clark (1987). A new methodology for modelling failure patterns in deteriorating water distribution systems: Theory. *Adv. Water Resources*, **10**, 2–10.
- Asbeck, E. and Y. Y. Haimes (1984). The partitioned multiobjective risk method. *Large Scale Syst.*, **6**, 13–38.
- Chankong, V. and Y. Y. Haimes (1983). *Multiobjective Decision Making: Theory and Methodology*. Elsevier-North Holland, New York.
- Haimes, Y. Y. and W. A. Hall (1974). Multiobjectives in water resources systems analysis: The surrogate worth trade-off method. *Water Resources Res.*, **10**, 615–624.
- Haimes, Y. Y. and D. Li (1988). Hierarchical multiobjective analysis for large-scale systems: Review and current status. *Automatica*, **24**, 53–69.
- Haimes, Y. Y., D. Li and E. Z. Stakhiv (1990a). Selection of optimal flood warning threshold. In Haimes, Y. Y. and E. Z. Stakhiv (Eds), *Risk-Based Decision Making in Water Resources*. American Society of Civil Engineering, New York.
- Haimes, Y. Y., K. Tarvainen, T. Shima and J. Thadathil (1990b). *Hierarchical-Multiobjective Analysis of Large-Scale Systems*. Hemisphere, New York.
- Kahmaman, D., P. Slovic and A. Tversky (Eds) (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, U.K.
- Karlsson, P. and Y. Y. Haimes (1988a). Risk-based analysis of extreme events. *Water Resources Res.*, **24**, 9–20.
- Karlsson, P. and Y. Y. Haimes (1988b). Probability distributions and their partitioning. *Water Resources Res.*, **24**, 21–29.
- Karlsson, P. and Y. Y. Haimes (1989). Risk assessment of extreme events: Application. *J. Water Resources Planning and Management*, **15**, 299–320.
- Li, D. and Y. Y. Haimes (1987a). A hierarchical generating method for large scale multiobjective systems. *J. Optimiz. Theory Applic.*, **54**, 303–333.
- Li, D. and Y. Y. Haimes (1987b). The envelope approach for multiobjective optimization problems. *IEEE Trans. Syst. Man Cybern.*, **SMC-17**, 1026–1038.
- Li, D. and Y. Y. Haimes (1987c). Risk management in a hierarchical multiobjective framework. In Sawaragi, Y., K. Inoue and H. Nakayama (Eds), *Toward Interactive and Intelligent Decision Support Systems*, Vol. 2. Springer, Berlin.
- Li, D. and Y. Y. Haimes (1988). Decomposition technique in multiobjective discrete-time dynamic problems. In Leonides, C. T. (Ed.), *Control and Dynamic Systems*, Vol. 28. Academic Press, San Diego, CA.
- Li, D. and Y. Y. Haimes (1990). Multiobjective control of risk of extreme events in dynamic systems. Presented at the IXth International MCDM Conference, Washington, D.C.
- Lowrance, W. W. (1976). *Of Acceptable Risk*. William Kaufmann, Los Altos, CA.
- Mays, L. W. and M. J. Cullinane (1986). A review and evaluation of reliability concepts for design of water distribution system. Report to Department of the Army, U.S. Army Corps of Engineers, EL-86-1.
- Mitsiopoulos, J. and Y. Y. Haimes (1989). Generalized quantification of risk associated with extreme events. *Risk Analysis*, **9**, 243–254.
- Travainen, K. and Y. Y. Haimes (1982). Coordination of hierarchical multiobjective systems: Theory and methodology. *IEEE-SMC*, **12**, 751–764.
- U.A. National Academy of Engineering (1988). *Cities and Their Vital Systems: Infrastructure, Past, Present, and Future*. National Academy Press, Washington, D.C.
- U.S. Nuclear Regulatory Commission (1981). *Fault Tree Handbook*. NUREG-0492.

Expert Systems—Principles and Programming*

Joseph C. Giarratano and Gary Riley

Reviewer: R. JIROUŠEK

Czechoslovak Academy of Sciences, Institute of Information
Theory and Automation, Pod vodarenskou veží 4, CS 182 08
Prague 8, Czechoslovakia.

EXPERT SYSTEMS—one of the most frequent terms in many fields of present human professional activities—as a term provoking a lot of questions and contradictory opinions. On the one hand, the motto “the problem is too hard to cope with, we need an expert system” has penetrated the subconscious of many specialists and managers-in-chief, on the other, there are still papers published, such as Streithberg (1988) stating the nonexistence of “real” expert systems. Though there is no exact definition of the expert system unanimously accepted by the Artificial Intelligence community, the notion has gained its indisputable place in AI and has become a regular part of AI courses for both undergraduate and graduate students.

The book by Joseph Giarratano and Gary Riley is the only one I know which aims to serve as a basis for courses on expert systems supplemented with a term project enabling students to develop skills in their design and programming. For this purpose, the second part of the book presents a detailed account of an expert system programming language CLIPS and the book is supplemented with disc containing a complete, executable expert system tool integrated with the text.

CLIPS (C Language Integrated Production System) was designed and developed at the Artificial Intelligence Section of the NASA/Johnson Space Center to facilitate laborious programming work connected with building expert systems. Regarding the differences between writing a program in a procedural language and implementing an expert system, it is very important for students to complete one or two small expert systems of their choice. Their task should not be only the implementation of the given techniques but the project should consist primarily of choosing the most appropriate method for the problem in question. For this purpose, a wide range of methods and techniques used for knowledge representation and processing in expert systems constitutes the content of the first part of the book.

As already mentioned above, there is no exact definition of the expert system. To be called an expert system, a program product has to meet several requirements. There is a common consensus about some of them; others are claimed only by some authors. And yet, it is very difficult, almost impossible, to say that some requirements are more important than others. All this wavering is reflected quite naturally in the first part of the introductory Chapter 1. Though there is only one paragraph entitled “What is an expert system”, it takes, in fact, nine of them to answer the question. The remainder of Chapter 1 is devoted to explanation of the most popular techniques and paradigms employed in expert systems today. In addition, as an example of a prospective technique, *artificial neural systems* are introduced.

At this place I would like to stress one very positive feature of the whole book. Whenever a new term is mentioned, it is illustrated with examples and relations to other similar concepts are explained. When, for example, *production systems* are introduced, the authors do not restrict themselves only to stating that knowledge is represented in a form of IF-THEN rules, and go back to the history and

show the origins of the production (rule-based) systems in *Post production systems* and *Markov algorithms*.

The same pedagogical way of presentation is used also in further parts of the book that expand into a detailed discussion on the theory (or rather theories) behind expert systems. The theory explains advantages and disadvantages of individual methods and therefore its knowledge is necessary when designing an expert system for a specific field of application. Chapter 2 describes different ways of knowledge representation. Though the most popular technique is based on productions, the authors bring in also *semantic nets* and *frames*. They do not explain only scopes and advantages of these methods but also difficulties with their application (a paragraph describing difficulties with production systems is unfortunately missing). A substantial part of Chapter 2 is also devoted to logic: *propositional logic* and *first order predicate logic*. This creates a necessary theoretical background for Chapter 3 dealing with methods of inference. In this chapter, *resolution*, the primary inference mechanism of PROLOG, is explained. But naturally, in accordance with other chapters, the authors pay attention to several other methods like *deduction*, *induction*, *abduction* or *nonmonotonic ways of inference*. All these methods are again described on examples and compared with one another.

Accepting the authors' statement that “in the real world we are seldom absolutely sure of anything except death and taxes” (and I believe this fact is one of the main impulses leading to the origin of expert systems), one has to consider Chapters 4 and 5 to be among the most important theoretical parts of the book. They form a wide survey of methods designed for both reasoning under uncertainty and inexact reasoning. In contrast with logical approach described in the previous chapters, these methods are important for expert system application involving uncertain information. The uncertainty may apply to facts (data), knowledge (rules) or both. A number of techniques has been suggested to deal with uncertainty in expert system but some of them are heuristic, not supported by a mathematical theory. It is a virtue of the book that it emphasizes theoretically sound approaches based on either *classical probability theory*, *Dempster-Shafer theory* or *Zadeh's fuzzy theory*. All three approaches are introduced in the book with help of examples some of which are used several times to show the differences between the individual models.

I have hitherto commented mainly positive aspects of the book. Now, I would like to mention a certain imperfection affecting presentation of some theoretical parts and issuing from the extensive choice of theories included in the book.

For example, when the classical probability theory is presented, basic terms like *sample space*, *random variable*, *probability distribution*, *conditional probability* or *independent events* are explained and illustrated with examples. Nevertheless, the authors do not introduce concepts like *multidimensional distribution* which would make possible to define the *dependence between random variables*—the notion used to represent knowledge in most of the probabilistic expert systems. Similarly, when discussing resolution as the principal inference rule in PROLOG it might be useful to mention its ways of implementation and connected problems (though it can be found quite easily almost in any PROLOG manual).

With a limited room for introduction of theoretical foundations of each described method it is fairly difficult to determine what can be omitted. To offset the subjectivity of their decision, the authors have supplemented each chapter with bibliography. It enables the lecturers, who take the

* *Expert Systems—Principles and Programming* by J. C. Giarratano and G. Riley. P.W.S.-KENT, Boston (1989). ISBN 0878353356, £16.95.

book as a basis for their courses, to go into more detail in the parts of their choice and to skip some others. There is no doubt that from this point of view the book is a success. This feeling is increased also by the fact that each chapter is supplemented with exercises and problems. Nevertheless, the book taken as the only source of information may leave a reader with a rather superficial idea about theory of expert system. Let us stress that this fact does not, however, decrease the value of the book as the basis for lectures on expert system theory and programming. For this reason, I would like to repeat the words from L. A. Zadeh's foreword: "Joseph Giarratano and Gary Riley deserve the thanks of all of us for undertaking a difficult task and making an important contribution to a better understanding of the fundamentals of expert systems".

Reference

Streitberg, B. (1988). On the nonexistence of expert systems—Critical remarks on artificial intelligence in

statistics. *Statistical Software Newsletter* 19, 2, 55–62.

About the reviewer

Radim Jiroušek was born in Prague, Czechoslovakia. He graduated in mathematics from Charles University, Prague, in 1969, and received the CSc (an equivalent of Ph.D.) degree in theoretical cybernetics from the Czechoslovak Academy of Sciences in 1979.

In 1970 he joined the Medical Cybernetics Laboratory at the Institute of Hematology and Blood Transfusion, Prague, where he researched mathematical methods of diagnosis. In 1979 he joined the Information Theory Laboratory, Institute of Information Theory and Automation of the Czechoslovak Academy of Sciences, Prague, where he is head of the Department of Decision Systems Theory. His present interest lies in uncertainty processing in expert systems with emphasis on probabilistic methods. He has published over 40 scientific papers, and at present is scientific secretary of the Czechoslovak Cybernetic Society.

Large Space Structures: Dynamics and Control*

S. N. Atluri and A. K. Amos

Reviewer: E. GOTTZEIN

MBB/Deutsche Aerospace, Space Communications and Propulsion Systems Division, Messerschmitt-Bölkow-Blohm GmbH, D 8000 München 80, Germany

AUTOMATIC control of large flexible space structures is one of the key issues in future spacecraft design. The problem is characterized by control-structure interaction resulting from a high number of densely packed structural modes within the control system bandwidth or close to it, combined with stringent micro-gravity level or pointing performance requirements. Examples for control systems of the first type are in-orbit manufacturing facilities with large solar panels and of the second type are very large space telescopes and laser pointing systems requiring pointing and target tracking to an accuracy of 0.01 arcsec or better. The problem to be solved is therefore not how to control or stabilize some large structure somehow but rather how to design both structure and control system in order to meet stringent system engineering requirements in an optimal way. This has to be kept in mind in order to frame the right question and solve the real problem.

For this new type of space systems, control engineers and mechanical engineers have to come together early on a system engineering level to properly match control systems and structural design parameters to meet overall requirements, and to predict on orbit performance of the space platforms as closely as possible. These two steps, control system design and overall performance evaluation, require different types of mathematical models—in particular those of the structure: a design model approximating the essential features of the system and a validation or truth model approximating the real system as closely as possible.

Mathematical models used in control system design have to properly represent transfer functions and responses between sensors and actuators, e.g. those located on different points of a structure. It is on this requirement that the mathematical modelling of the structure for control system design has been based. Approximations which are quite different from the ones which are commonly used in structure design may lead to better and less complex control

figurations. Even the extreme case has been demonstrated, where control problems which were unsolvable or only solvable at the expense of large complexity became solvable with technically feasible controllers by re-evaluating the modelling approach. The iterative procedure between control system design and modelling approximation is therefore the first key issue in control-structure interaction.

The second key issue is how to design robust controllers to stringent performance requirements under high levels of uncertainty in the mathematical model itself due to the finite approximations used, the model parameters and the disturbing environment. Robustness is required to assume stability and minimum in-orbit performance in the first place while fine tuning can be done later by in-orbit identification and adaptation of control configuration and parameters with the spacecraft already flying in its real environment.

A third key issue is verification of mathematical models and closed loop system performance by cleverly designed experiments on ground and in orbit. Even with the powerful design tools available nowadays, the importance of doing the right tests on a component as well as on a control system level cannot be overemphasized. Only by testing can it be proven that the control system design problems have been solved correctly and that the right problems have been solved. Concerning dynamics and control of large flexible space structures, we are, with all three key issues, just in the beginning.

The monograph on dynamics and control of large flexible space structures addresses the first two key issues. Its fifteen chapters which are written by structural dynamicists and control "theoreticians" are intended to give the status of this highly interdisciplinary subject and an indication of future trends. Stimulating early research work is reported; unfortunately the chapters stand pretty much by themselves and are only very loosely related to each other. They are, in general, well written and give adequate references. In accordance with the status given in the monograph, the desired interaction between structural and control specialists is not yet achieved. The chapters on structural modelling on one hand oversimplify the control problem, and the chapters on control system design, on the other hand, are based on linearized modal characteristics (an approach which is well supported by practical experience so far) but do not question the assumptions. The weakness of the monograph is that the underlying control performance requirements, which are the real cause of the problem, are hardly stated at all and that

* *Large Space Structures: Dynamics and Control* by S. N. Atluri and A. K. Amos. Springer, New York (1988). ISBN 0-387-18900-9, U.S. \$89.50.

too few realistic examples are given to demonstrate the applicability of the described procedures although these examples are surely available (Auburn and Lorell, 1987; Elbuni and Higashiguchi, 1989).

Fortunately there are exceptions: In the chapter on "Modal Cost Analysis for Simple Continua" Hu, Skelton and Yang use modal cost analysis as a tool for selection of finite element models and base the necessary modal reduction decisions on control objective functions. The "Model-Control Inseparability Principle" is demonstrated on the particular example of an Euler-Bernoulli-beam. Convergence of modal costs and of the closed loop design to a satisfactory level is better using quintic beam elements instead of the commonly used cubic ones. The physical background behind this is that in control system design, accuracy of modal shapes and slopes is more important than accuracy of modal frequencies.

The limits of finite dimensional approximations and modal representations are discussed in two chapters. Zak in "Dynamic Response to Pulse Excitations in Large Space Structures" investigates stress pulse propagation along repetitive lattice type structures including reflections and transitions at joints. He suggests modelling by system equations with delay arguments. Von Flotow in "The Acoustic Limit of Control of Structural Dynamics" distinguishes between modal analysis with a global description of entire structure and wave propagation analysis based on scattering properties of local components. He defines the acoustic limit of structure dynamic modelling by the limitation of modal analysis which is the sensitivity of the densely packed high frequency modes to modelling errors. Beyond this limit, modal equations are useless for control system design. Alternatives are active control of lower modes and passive damping or local active structural control based on acoustic models. This is described as "travelling wave control". Various examples for travelling wave control are given including experimental verification.

In "Continuum Modelling of Large Lattice Structures", Noore and Mikulas develop equivalent continuum models for large repetitive truss and frame lattice structures of the beam or shallow shell type. Numerical examples are evaluated on the basis of frequency comparison with finite element approximations.

In "Nonlinearities in the Dynamics and Control of Space Structures" Atluri and Iura deal with nonlinearities stemming from large deformations and rotations of the structures and their members, changes in the inertia matrix due to these deformations, damping due to nonlinear hysteresis and flexible joints and nonlinear behaviour of the structural material. Tangent stiffness operators and tangent inertia operators are defined for piecewise linearization. The concept of active control or damping by piezo-ceramic actuators that are bonded to the truss and frame members in various locations is considered in detail.

Friction forces at structural joints and interfaces are possible elements for energy dissipation and passive damping but may also be the cause of limit cycles in closed loop systems. In "Dynamic Friction" Srinivasan develops refined models to describe friction between contacting surfaces. The investigations are backed by numerous experiments.

Modi and Ibrahim develop in "On the Transient Dynamics of Flexible Orbiting Structures" relatively general formulations to describe arbitrary multi-body systems where each body may have an arbitrary elastic structure. The idea is to develop a comprehensive data bank for spacecraft with flexible appendages. The formulations are applied to the orbiter with deployable flexible appendages.

In "A Review of Modelling Techniques for the Open and Closed-Loop Dynamics of Large Space Structures" Bainum compares various approaches to structural modelling and control system design.

In "Computational Issues in Control-Structure Interaction Analysis" Park deals with general solution procedures for dynamical problems. The relevance to control-structure interaction is not shown and the examples given are from outside the area.

Meirovitch in "Control of Distributed Structures" bases

his control system design solely on modal techniques. It is unfortunate that no numerical examples what so ever are given to show the feasibility of the conclusions.

There are only three chapters on control system design in the stricter sense focused on how to achieve robust control under uncertainty in model equations, model parameters and system disturbances.

Lynch and Banda in "Active Control for Vibration Damping" use "linear quadratic Gaussian with loop transfer recovery (LQG/LTR) design technique" to develop a robust vibration control system. LQG/LTR modifies the conventional LQG problem into a loop shaping problem which is dealt with through the singular values of the open loop transmission matrix. The advantage of the method is that control objectives can intuitively be represented as bounds in the frequency domains and that it allows for establishment of uncertainty profiles based on the difference between the design and truth model. In the numerical example given, an active vibration damping control system is designed using LQG/LTR technique for a two bay truss structure. It is demonstrated that robustness recovery has to be balanced versus actuation power and that the control power requirements set the limit on how closely a desired open loop shape may be approximated.

Bernstein and Hyland in "Optimal Projection for Uncertain Systems" extend the standard LQG to cover reduced order constraints on the dimension of the dynamic compensator and uncertainty modelling of deterministic parameter uncertainties and stochastic disturbances. Some numerical results are given to compare the optimal projection approach to various LQG reduction techniques. In this comparison optimal projection scores very well. Unfortunately the authors describe the underlying problems only vaguely and refer the reader to the references.

Kosut in "Adaptive Control of Large Space Structures" uses adaptive uncertainty modelling to obtain frequency domain expressions for characterizing a "set of uncertainty" within which the true plant lies. It is suggested that the adaptive process, referred to as adaptive calibration is used to extract the set of uncertainty from online measurements thus providing an approach to in-orbit redesign and fine tuning of the control system. Parametric system identification is used to obtain a nominal estimate of the plant transfer function. Nonparametric spectral estimation methods are used to obtain a frequency domain expression for the model uncertainty of the nominal estimate. The resulting controller is designed to be robust with respect to the estimated set of uncertainty. A large set of uncertainty will lead to a low authority controller and a small set of uncertainty to a high authority controller. The uncertainty estimation procedure is applied to a laser pointing and control experiment, giving promising results.

The last two chapters are devoted to simultaneous structure and controller design optimization. They prepare the way for trade-offs of controller versus structure complexity. Parameters to be optimized are structure weight, control power consumption, stiffness of members, etc.

In "Unified Optimization of Structures and Controllers" Junkins and Rew start out from the "Eigenstructure Assignment Theorem" and discuss two recently developed generalizations of the quadratic regulators for robust eigenstructure assignment. They exchange the nonlinear algebraic Riccati Equation by the simpler linear algebraic Lyapunov Equation in the determination of the optimal gain matrix. In the numerical example of structure versus controller design a two body structure consisting of a rigid body with a flexible beam attached by a torsional spring is considered. The problem is solved by sequential linear programming techniques. The different design objectives considered are minimum eigenvalue sensitivity, maximum stability robustness and minimum mass. They lead to distinctly different solutions for actuator locations, structural parameters and optimal control gains and impressive mass savings. The chapter is theoretically interesting and ambitious so that one would really like to see more than simplified examples.

In "An Integrated Approach to the Minimum Weight and

Optimal Control Design of Space Structures Khot considers two approaches to minimize structural weight, under constraints on closed loop eigenvalues and control gains' norm.

Unfortunately the underlying examples are different and allow no comparison of results.

Modelling and control of large flexible space structure is presently investigated and applied in many places and one can hardly expect one book to give a full account. Of the missing subjects, I want to mention at least a few of engineering importance. For setting up mathematical models of complex structures, the Craig-Bampton Component Model Analysis is to be mentioned. On the control side high authority control/low authority control (HAC/LAC) is barely mentioned, One Controller at a Time (1-CAT), Filter Accommodated Model Errors Suppression (FAMES) and Positivity are not mentioned at all. These methods have been widely studied and even tested in various research programs (Mitchell *et al.*, 1984, 1985).

The monograph is not a textbook for learning. It addresses the researcher in the fields of Dynamics and Control of Large Space Structures, and for those specialists it is highly recommended. As a practising engineer in spacecraft system design, I enjoyed very much reading it: My adrenalin level was continuously high either from excitement or criticism. The monograph is restricted to the theoretical treatment of problems only. The third key issue verification and test is not addressed. The reader of this review might be interested to learn that the Aerospace Committee of IFAC held a workshop on "Modelling and Validation of Flexible Aerospace Structures and Aerospace Control" in Huntsville, Alabama in April 1991.

References

- Auburn, J.-N. and K. R. Lorell (1987). Performance analysis of the segment alignment control system for the ten meter telescope. *10th IFAC World Congr.* Munich, pp. 181-191.
- Elbuni, M. S. and M. Higashiguchi (1989). Active stabilization of a large flexible antenna feed support structure. *11th IFAC Symp. on Automatic Control in Aerospace.* Tsukuba, Japan.
- Mitchell, J. R., S. M. Seltzer and D. K. Tollison (1984). 1-CAT (One-controller-at-a-time): A frequency domain multi-input multi-output design approach. *Proc. 1984 AIAA Guidance and Control Conf.*, Seattle.
- Mitchell, J. R., S. M. Seltzer and H. E. Worley (1985). Design-to-performance. *Proc. 1985 IFAC Conf. on Automatic Controls in Space.* Toulouse, France.

About the reviewer

Eveline Gottzein studied electrical engineering, applied mathematics and control and mechanical engineering at the Technical Universities of Dresden, Darmstadt and Munich. She received the degree of Doktor-Ingenieur at the TU Munich. After various employments in process control and simulation of chemical and nuclear plants, she joined MBB GmbH in 1959. Her present position is head of the Control Dynamics and Simulation department. Her activities are in the fields of guidance, navigation and control of spacecraft such as launch and reentry vehicles and satellites. Since 1974, she has been a member of the TC on Aerospace in IFAC and served several terms as chair/vice-chairperson. She lectures at the Technical University of Stuttgart on the subject of Control of Aerospace Vehicles.

Applied Control of Manipulation Robots— Analysis, Synthesis and Exercises*

Miomir Vukobratovic and Dragan Stokic

Applied Dynamics of Manipulation Robots— Modelling, Analysis and Examples†

Miomir Vukobratovic

Reviewer: T. VAMOS

Computer and Automation Institute, 1133 Budapest, Victor Hugo u. 18-22, Hungary.

AFTER A successful series of six volumes on robotics, we now receive two further ones authored by the same person and school: Miomir Vukobratovic, a professor working with the Mihajlo Pupin Institute in Belgrade, Yugoslavia. The two new volumes are devoted to the dynamics of manipulation robots (i.e. on the most widespread type of robots) and on their control. They are textbooks of undergraduate and graduate courses, a thorough guide for learners of modelling, analysis, synthesis and design. The course is defined by this

* *Applied Control of Manipulation Robots—Analysis, Synthesis and Exercises* by M. Vukobratovic and D. Stokic. Springer, Berlin (1989) 470 pp., 100 figures. ISBN 3-540-51469-4, DM 138.00

† *Applied Dynamics of Manipulation Robots—Modelling, Analysis and Examples* by M. Vukobratovic. Springer, Berlin (1989), 471 pp., 176 figures. ISBN 3-540-51468-6 DM 138.00.

objective; it provides as few detours for general knowledge or information about the broad subject as possible, and as such is not a usual review of robotics.

This goal determines both the values of the approach and its limitations. The main thrust of both volumes is the computation procedure of robot design and therefore all detailed calculations, equations with multidegree of freedom, multi-joint constructs are given with a very practical addition of the corresponding computer programs written in Fortran-77.

In the first volume, an introductory part deals with robots in general and their classification. This introduction, however, more serves the further chapters than gives the usual insight into techniques and application, as it is specially oriented to the understanding of the further text, i.e. the requirements of the different degrees of freedom for manipulation.

The thrust of the first volume is in the second chapter, and extended over three quarters of the basic text, more than the half of the appendices, and references. It provides the complete mathematical model of the manipulator robots and

their dynamics. Special attention is given to vibration, flexible manipulation and coordinated manipulation problems. These are extremely delicate calculation tasks; deformation and vibration are propagated by the joints to further parts of the robot, having different kinds of degrees of freedom, and most of the results are originated by the author himself or by his school under his scientific leadership. These delicacies are mostly neglected in the usual textbooks, handbooks and general reviews of the subject, yet they play a relevant role in practice. Several times they are the decisive factors of appropriate selection of the robot and of detection of the limits of its usage. The third chapter of Vol. I is virtually a research report by the author on linearization and sensitivity calculations which lead to the computer models of the design. A surprising arrangement of the book is this volume's nine appendices, constituting about the half of the full text. This is understandable from the objective of the course; the reader should get a detailed practice in these very specific calculations, and those who complete all these should master the method, and not only know about it in general.

The same refers to the other volume. Somebody who has an expertise in dynamics should continue in the same way in control. In a rather similar way to the first volume, the first chapter gives a general overview of robot control, the second chapter that of the kinematic level of control with a special attention to its basic problem, the inverse kinematic and the synthesis of the trajectory based on the same. Here the third chapter contains the main thrust: the synthesis of servo systems and how the individual control elements should be designed for the fulfillment of the overall manipulation task. This is synthesized in chapter 4 dealing with the simultaneous motion control with a special attention to the stability and the possibility of a decentralized control. The global control follows in chapter 5, variable parameters and adaptivity in chapter 6 and the considerations of motion constraints in chapter 7. The construction of the second volume is slightly different from the first, the appendices follow immediately the chapters partly for exercises, partly for amendments for some more delicate details. The slight differences in the style of the two volumes is due to an associate author, Dragan Stokic of the same group, but also to the specific research interests of the authors which unavoidably shift the proportions of treatment in the direction of those details which are closer to their personal scope of activity. The authors offer the computer programs of these calculations as software products with documentation prepared for IBM-PCs.

Who should be the reader of these books? Less application people than those who intend to design the mechanics and control of manipulators. However, the course is very useful for anybody who will go further in his/her life in the design of any moving mechanical systems and their control. Such a thorough exercise, detailed analysis, synthesis practice is a lifetime experience of what should be considered in the

design of a complex, interconnected, dynamic system, how should the controls of individual components and tasks should be coordinated, and how many viewpoints should be compromised. In educating engineers the school cannot prepare the young generation for all the tasks they will encounter in their lives—technology, economy, changing opportunities—but a disciplinary training, an internal drive for a professional approach, a hard lesson of a certain task completion is a legacy for a whole professional life. What would I like to see somehow different? Firstly, a more didactic introduction to each topic, starting with the basic physics of the equations and an approach by coming down from a bird's-eye view to the specific equations. The authors can answer that this is done by the related courses; but in a textbook one would like to see more reference to the required basic knowledge and viewpoint. A third volume, *Introduction to Robotics*, may serve this general introduction but a guide for the full course, how to divide it and how to have a logical continuation would have been preferred, with some pedagogical hints based on the experience of the authors. Similarly, all details of practical realization are omitted, such as technological problems, economy, etc. These are beyond the objective of the authors and the reviewer has to be resigned to it; however, the student and the design engineer must look at the task in such a totality; this is just the engineering view. Omissions are unavoidable if the authors wanted to give such a detailed guidance to the design of these problems; they could not get into all the topics of a general survey. On the one hand this specific goal and its thorough fulfillment is the virtue of the books, but on the other, at least the interfaces to others should be given.

These are minor remarks from a differing viewpoint, and do not refer to the value of the unique course. The series and its predecessors are unique undertakings and therefore not only students of robotics and kinematics but all people involved in robot design should have them on their shelves.

About the reviewer

Professor T. Vámos was born in Budapest in 1926, and graduated from the Technical University of Budapest. He initially studied process control automation of power plants and systems, later computer control of processes, has been working for 10 years in the industrial pattern recognition-robot vision field, and is now engaged in expert systems. He is professor of the Technical University of Budapest; Chairman of the Board of the Computer and Automation Institute, Hungarian Academy of Sciences, Past President of the International Federation of Automatic Control; Fellow of the IEEE; Member and Board Member of the Hungarian Academy of Sciences, Doctor *Honoris Causa* of the Tallinn Technical University; and Honorary President of the John v. Neumann Society for Computing Sciences. Prof. Vámos is the author of more than 120 publications and invited speaker of several international congresses and conferences.

Dynamic Models and Discrete Event Simulation*

William Delaney and Erminia Vaccari

Reviewer: I. M. Y. MAREELS

Department of Electrical and Computer Engineering, The University of Newcastle, Rankin Drive, Newcastle, NSW 2308, Australia.

* *Dynamic Models and Discrete Event Simulation*, by W. Delaney and E. Vaccari. Marcel Dekker, New York (1989). ISBN 0-8247-7654-2, U.S. \$99.75 (U.S.), \$119.50 (other countries).

THE BASIC aim of the authors is to provide a wide public of undergraduates with a systematic approach to modelling and subsequent simulating of real life phenomena (systems). The model is perceived as an abstract representation of a system allowing reproduction of the system's behaviour (to varying degrees of completeness) aiming at elucidating one's understanding of the system and predicting its future behaviour. The book is not about control. The emphasis is to clarify the limitations and potentials of models at least from a

conceptual point of view, without relying on high level technical/mathematical development.

The first half of the book is devoted to the fundamental system/information theory problem: "From raw data to model". Model identification issues are discussed for a large variety of models: continuous time/discrete time, deterministic/stochastic, lumped parameter/distributed systems, time invariant/non-stationary, autonomous/forced, linear/nonlinear.

The second half is more concerned with software engineering issues related to simulating, i.e. algorithmically approximating the behaviour of the model on a digital computer. The important application of simulation stressed here is its role in guiding experiment design and data conditioning for model identification and subsequent model validation.

Examples and exercises are amply provided. Some simulation studies are presented in great detail using the BDSIM general purpose simulation package developed by the authors.

The mathematical and computer science background required to access the material is kept to the bare minimum. An elementary knowledge of linear algebra and calculus together with some Fortran (?) literacy suffices to understand the technical content of the book. The material is intended to be easily accessible by any science or engineering student in the second year of the curriculum.

As far as I am aware the book is unique in its kind. It is to the authors' credit to have tried to accomplish the above daunting educational task. Most available textbooks cover either the system theoretic aspects or the software engineering (simulation oriented) issues and none require so little prior technical knowledge of the reader. Moreover most (undergraduate) textbooks dealing with system theory are only concerned with very limited classes of systems (typically those for which control theory is well developed) and books about linear systems abound. Nonlinear systems are less popular and discrete event systems are largely forgotten. Good graduate level textbooks or reference texts dealing with modelling identification (at least for linear systems) are available (e.g. Caines, 1988; Ljung, 87). An excellent series of articles by J. C. Willems (Willems, 1986) can also be recommended. Discrete event systems seems to be more the realm of simulation oriented textbooks (e.g. Bulgren, 1982; Fishman, 1978).

Unfortunately, the authors have opted for breadth rather than depth. Although they state "to avoid superficiality, concepts and logic development are strongly emphasized", the book gives a rather confused, blurred view on both modelling and simulation (falsely called a system theoretic approach). Moreover, the interested reader, who wants to read more to obtain a clearer picture, is not referred to the correct literature, at least as far as the system theoretic aspects are concerned (none of the above references are quoted).

The basic notion of system is defined in Chapter 1 on page 1 as "a set of subsystems, related so as to form a unit". The following 50 pages fail to elucidate this "definition". In the same chapter concepts as environment ("influencing but not being influenced"), inputs, outputs, states, behaviour, causality are equally confusedly introduced. About 40 "fundamental" concepts are proposed without proper definition. If the student is not lost at this point, and not completely disgusted with the apparent imprecision of "system theory/philosophy", there is more . . .

The next three chapters discuss deterministic, and stochastic models rather by example. The exposition is directed towards identification of these systems. The treatment is extremely superficial, and, therefore, in my opinion, unintelligible. How can one hope to convey the basic principles for deterministic systems of state space and input-output methods, stability theory (using energy concepts), transfer functions, exact and approximate analytic and numerical solution methods in less than 60 pages? Even more superficial is the discussion of stochastic systems. Here

the reader is hurried from the basic definition of probability as given by B. Pascal to the Cramer-Rao bound for information content, in less than 200 pages. In the mean time the reader is exposed to linear (and nonlinear) regression algorithms in correlated noise environment applied to data generated by a possibly (mildly) nonlinear system! This is quite inconsistent with the educational intentions envisaged by the authors. Clarity and understanding are abandoned in favour of encyclopedic coverage of material.

Chapter 5 forms the bridge between the model based first half and the software engineering treatment of simulation of the sequel. It discusses complexity as a motivation for simulation studies. Tacitly it is assumed that it is actually possible to build a relevant model (e.g. from models of subsystems) for the complex system. Here and in the sequel the emphasis shifts towards discrete event systems.

The discussion of sound simulation techniques and the interaction between modelling and simulation is completely at the descriptive level. No quantitative criteria for quality of software, algorithms and simulation studies are discussed. Some generic information about artificial intelligence (artificial reasoning) and its role in a general simulation context is also provided.

As a minor point, yet quite annoying, is the rather unfortunate discussion of pseudorandom number generators using the Lehmer's congruent method, presented in Chapter 7. This method is sound but the implementation discussed is by no means acceptable! (See e.g. Park and Miller, 1988 or Fishman, 1978).

From an educational point of view it is unfortunate that the book uses Fortran as its programming environment (with exception of the artificial intelligence discussion, where Prolog and Lisp are mentioned). Moreover no mention is made of the many software packages available for simulating (continuous event) systems such as Matrixx, Matlab, Easy5 . . .

In conclusion, it is a pity that the goals set by the authors themselves are not met in their book. Surely education should in the first instance strive for clarity, not quantity. A book introducing (at the undergraduate level) modelling and simulation in a unified framework, honouring this principle would be most welcome indeed!

References

- Bulgren, W. G. (1982). *Discrete System Simulation*. Prentice-Hall, Englewood Cliffs, NJ.
- Caines, P. (1988). *Linear Stochastic Systems*. Wiley series in Probability and Mathematical Statistics, Wiley, Chichester.
- Fishman, G. S. (1978). *Principles of Discrete Event Simulation*. Wiley, New York.
- Ljung, L. (1987). *System Identification, Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ.
- Park, S. and K. Miller (1988). Random number generators: good ones are hard to find. *Commun. ACM*, **31**, 1192-1201.
- Willems, J. C. (1986). From time series to linear systems. *Automatica*, **22**, 561-580, 675-694, 87-115. See also *Models for Dynamics*, (1988). **2**, 171-269.

About the reviewer

Dr I. Mareels obtained a B.E. in electro-mechanical engineering from the State University of Gent, Belgium in 1982. He received the Ph.D. degree in systems engineering from the Australian National University, Canberra, Australia in 1987. He is a member of SIAM and IEEE and is currently senior lecturer at the Electrical Engineering Department, Newcastle, Australia. His main research focuses on dynamical aspects of systems in signal processing, adaptive control, adaptive identification, power systems and nonlinear control theory.

Dr Mareels is an Associate Editor for *Automatica*, in the area of adaptive control and identification.

Biographical Notes



Keiichi Akimoto was born in Osaka City, Japan on 20 October 1951. He received the M. E. in engineering of applied mathematics and physics from Kyoto University in 1976, and joined the Kawasaki Steel Corporation. Since then, he has been engaged in the development of the measurement and instrumentation systems for the steelmaking processes. He is currently Staff Manager of the Electrical and Instrumentation Technology Section in the Kawasaki Steel Mizushima Works. He is a member of the Society of Instrument and Control Engineers.

He is currently Staff Manager of the Electrical and Instrumentation Technology Section in the Kawasaki Steel Mizushima Works. He is a member of the Society of Instrument and Control Engineers.



Michael Athans was born in Drama, Greece on 3 May 1937. He received his B.S.E.E. in 1958 (with highest honors), his M.S.E.E. in 1959, and Ph.D. in control in 1961, all from the University of California at Berkeley.

From 1961 to 1964 he worked at the MIT Lincoln Laboratory, Lexington, MA. Since 1964 he has been with the MIT Electrical

Engineering and Computer Science Department where he is Professor of Systems Science and Engineering. He also served as director of the MIT Laboratory for Information and Decision Systems (formerly the Electronic Systems Laboratory) from 1974 to 1981. He is co-founder of ALPHATECH, Burlington, MA, where he is Chief Scientific Consultant and Chairman of the Board of Directors, and has also consulted for other industrial organizations and government panels.

Dr Athans is co-author of *Optimal Control* (McGraw Hill, 1966), *Systems, Networks, and Computation: Basic Concepts* (McGraw Hill, 1972), and *Systems, Networks, and Computation: Multivariable Methods* (McGraw Hill, 1974). In 1974 he developed 65 TV lectures and accompanying study guides on *Modern Control Theory*. He has authored or co-authored over 250 technical papers and reports. His research interests span the areas of system, control, and estimation theory and its applications to the fields of defence, aerospace, transportation, power, manufacturing, economic and command, control, and communications (C³) systems.

He has received many awards including the American Automatic Control Council's 1964 Donald P. Eckman award for outstanding contributions to the field of automatic control; the 1969 F. E. Terman award of the American Society for Engineering Education as the outstanding young electrical engineering educator; and the 1980 Education Award of the AACC for his outstanding contributions and distinguished leadership in automatic control education. In 1973 he was elected Fellow of the IEEE and in 1977 Fellow of the American Association for the Advancement of Science. He has served on numerous committees of IFAC, AACC and IEEE; he was president of the IEEE Control Systems Society from 1972 to 1974. In addition he is a

member of Phi Beta Kappa, Eta Kappa Nu and Sigma Xi. He was associate editor of the *IEEE Transactions on Automatic Control*, co-editor of the *Journal of Dynamic Economics and Control* and associate editor of *Automatica*.



Joseph Bentsman was born in Minsk, U.S.S.R., on 16 February 1952. He received the M.S.E.E. degree from the Byelorussian Polytechnical Institute, Minsk, U.S.S.R., in 1979, and the Ph.D. degree in electrical engineering from the Illinois Institute of Technology, Chicago, in 1984.

From 1975 to 1980, he worked as an engineer in the Design Bureau of Broaching Machine Tools, Minsk. In 1985, he was a lecturer and a postdoctoral research fellow in the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. At present, he is an Associate Professor in the Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign. His current research interests are in control of distributed parameter systems, nonlinear dynamics, and self-tuning control. Joseph Bentsman is a recipient of the 1989 National Science Foundation Presidential Young Investigator award in dynamic systems and control.



W. Andrew Berger received his M.S.E.E. and Ph.D. degrees from Drexel University, Philadelphia, PA, in 1984 and 1988, respectively. Since then he has been Assistant Professor of Electrical Engineering at the University of Scranton, PA. His research interests are in the areas of Multivariable Systems and Digital Signal Processing. Dr Berger is a member of the IEEE.



Jean-Paul Béziat was born in Castres, France, on 8 November 1961. He graduated from University Paul Sabatier, Toulouse, France, with a M.S. degree in 1984, and a DEA in 1986. He received the doctoral degree in control engineering from the same university in 1989. Since 1989, he has been employed by C.I.S.I., a company specialized in computers and services, in

Toulouse. His main fields of interest include adaptive identification and control, optimization and control under constraints.



Alessandro Casavola was born in Florence, Italy, in 1958. He received the Dr. Eng. degree from the University of Florence, Italy, in 1986. Since 1987 he has been a Ph.D. student at the Systems and Computer Engineering Department of the University of Florence. His research activity is primarily concerned with the polynomial approach to the H_2 and H_∞ optimal control

theory. His current research interests include optimal control, adaptive control and robotics.



Alessandro De Luca was born in Rome, Italy, on 11 October 1957. He received the Laurea in electronic engineering in 1982. In 1984 and in 1987, he received respectively the Masters and Ph.D. degrees in systems engineering, both from the Università di Roma "La Sapienza". From 1985 to 1986 he was a Visiting Scholar at the Robotics Laboratory of the Rensselaer

Polytechnic Institute, Troy, NY. Since 1988 he has been a Researcher at the Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza". His research interests are in nonlinear control systems, with applications to induction motors and robotics. In this latter area he is involved in trajectory planning, modeling and control of manipulators with elastic joints and flexible links, optimization schemes for kinematic redundancy resolution, and hybrid force-velocity control. Dr. De Luca is a member of IEEE.



Wei-Liang Chen was born in Taiwan in 1962. He received the BS degree in mechanical engineering from the Taiwan National University, Taiwan, in 1984, and the MS and PhD degrees in mechanical engineering from the University of California, Los Angeles, in 1987 and 1990, respectively. Dr. Chen is now an Associate Professor of Mechanical Engineering at Kaohsiung Polytechnic Institute, Taiwan. His main research interests are in robust control, system identification and adaptive control.



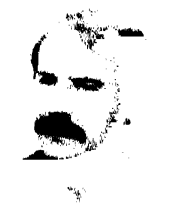
Jamel Fakhfakh was born on 15 July 1962 in Sfax, Tunisia. After graduating from high school with honors in 1982, he was awarded fellowships by the Tunisian Government to further pursue his undergraduate, and later, graduate education.

He received his B.S., M.S. and Ph.D. degrees in mechanical engineering in 1986, 1987 and 1990, respectively, all from the

University of Illinois at Urbana-Champaign. From 1986 to 1990 Dr. Fakhfakh served as a research assistant in the Department of Mechanical and Industrial Engineering and in the Energy Systems Division of USACERL. In 1990 Dr. Fakhfakh joined the Paulina Company in Tunisia. His interests are in control of distributed parameter systems, optimization methods, adaptive control, and nonlinear dynamics.



Dumas Chin graduated in March 1989 with First Class Honours in electrical engineering from the National University of Singapore. Upon graduation, he started employment with Singapore Aerospace as a Systems Engineer. He is currently with IBM Singapore.



J. Dastych is a scientific research assistant and senior lecturer at the Department of Electrical Engineering at Ruhr-University Bochum, Germany. Dr. Dastych studied electrical engineering and in 1983 submitted his doctoral thesis in control systems engineering. His present research activities have focused on identification and modeling of dynamic systems, industrial control applications,

simulation methods and tools, CAD systems for application in control systems engineering and robust control.



Gila Fruchter received the D.Sc. degree in mathematics from the Technion-Israel Institute of Technology, in 1988. She was then a Lady Davis Postdoctoral Fellow at the Technion. Presently, she is a visiting researcher in the University of California, Los Angeles. Her research interests are in robust control theory, where she has developed mathematical methods applicable for

stability and design problems, and in optimal control theory. Dr. Fruchter has been a member of IEEE since 1986.



J. S. Gibson received the BS in aerospace engineering, MS in engineering mechanics and PhD in engineering mechanics from the University of Texas at Austin in 1970, 1972 and 1975, respectively. He served on the faculties boards of the University of Texas at Austin and Virginia Tech before joining the faculty of UCLA in 1977. Currently, he is Professor of Mechanical, Aero-

space and Nuclear Engineering at UCLA. His research interests include control and approximation of distributed systems and adaptive control and identification.



Michael Grimbble was born in 1943 in Grimsby, England. He was awarded a first class BSc honours degree in electrical engineering from Rugby College of Technology in 1970. Subsequently he obtained MSc (1971), PhD (1974) and DSc (1982) degrees in control engineering from the University of Birmingham and a BA degree in mathematics from the Open University.

In 1971 he joined the Systems Engineering Department of GEC Electrical Projects Ltd. as a design engineer, and was promoted to Senior Engineer in 1974. He was then seconded to the Industrial Automation Group at the Imperial College of Science and Technology for research into Tandem Mill Automation.

The University of Strathclyde, Glasgow, appointed him to the Professorship of Industrial Systems in 1981, and he is now the Director of the Industrial Control Unit and Past Chairman of the Department of Electronic and Electrical Engineering. His group is concerned with industrial control problems, particularly those arising in the Aerospace, Wind Energy, Steel, Marine, Electricity and Gas supply industries. His research interests currently include self-tuning and *H_∞* robust control theory, multivariable design techniques, optimal control and estimation theory.

Professor Grimbble has chaired several IEEE and IFAC Working Groups and Societies. He is the Managing Editor of the *Journal of Adaptive Control and Signal Processing* and was an Associate Editor of *Automatica*. He edited the Prentice-Hall International series on Systems and Control Engineering and was the General Chairman of the IFAC Adaptive Control and Signal Processing Symposium held in Glasgow, 1989.

Professor Grimbble was awarded the IEEE Heaviside Premium in 1978 for his papers on control engineering and the Coopers Hill War Memorial Prize and Medal by the Institution of Electrical, Mechanical and Civil Engineering, in 1979.



Volker Hahn was born in 1954 in Dortmund, Germany. In 1978 he got the Dipl.-Ing. degree in electrical engineering from Ruhr-University in Bochum. From 1978 to 1984 he worked as a scientist at the institute of electrical control engineering of Ruhr-University Bochum with Prof. Dr.-Ing. H. Unbehauen. At that time he was concerned with adaptive control, especially direct adaptive control

for nonminimum-phase multivariable systems. In 1983 he got the degree "Dr.-Ing." In 1983 he joined the Uhde company as a process control engineer, and since 1985 he has worked as a control engineer in chemical industry, at Bayer AG.



Yacov Y. Haimes received the B.S. degree in 1964 from the Hebrew University, Jerusalem, Israel, the M.S. degree, in engineering in 1967, and the Ph.D. degree in engineering with distinction (majoring in large-scale systems engineering) in 1970, both from the University of California at Los Angeles.

He is Lawrence R. Quarles Professor of Engineering and Applied Science and Director of the Center for Risk Management of Engineering Systems at the University of Virginia, Charlottesville, Virginia. He is a former Chairman of the Department of Systems Engineering at Case Western Reserve University, Cleveland, OH. His research, teaching and consulting activities are in decision-making under risk and uncertainty and conflict resolution in large-scale systems within the hierarchical-multiojective framework, motivated by his extensive studies of water resources systems. From September 1977 to August 1978, he was the American Geophysical Union Congressional Science Fellow and spent three months in the Executive Office of the President and eight months with the U.S. Congress. He is the author/co-author of five books, the most recent of which is *Hierarchical-Multiojective Analysis of Large-Scale Systems*, published by Hemisphere Publishing Company, 1990. He is also the editor/co-editor of fourteen other volumes, the author/co-author of over 100 technical papers, Fellow of the IEEE and other societies, Vice Chairman of the IFAC Systems Engineering Committee (1987-90) and former Chairman of the Committee's Working Group on Water Resources Systems. He is an Associate Editor of *Automatica*, *IEEE Transactions on Systems, Man and Cybernetics*, *Control Theory and Advanced Technology*, *Information and Decision Technologies*, and *Reliability Engineering and Systems Safety*.



C. C. Hang graduated with First Class Honours in electrical engineering from the University of Singapore in 1970. He received the Ph.D. degree in control engineering from the University of Warwick, U.K., in 1973. From 1974 to 1977, he worked as a Computer and Systems Technologist in the Shell Eastern Petroleum Company (Singapore) and the Shell International Petro-

leum Company (The Hague). Since 1977, he has been with the Department of Electrical Engineering, National University of Singapore, where he is currently a Professor and Head of the Electrical Engineering Department. He was a Visiting Scientist in Yale University and Lund Institute of Technology in 1983 and 1987, respectively. His current research interests include computer process control, adaptive control and expert systems applications.



Jean-Claude Hennet was born in Tunis, Tunisia, on 28 November 1950. He received the diploma in aerospace engineering from Sup'Aéro, France, 1974, the M.S. Degree from Stanford University, 1975, the Docteur-Ingénieur diploma in Toulouse, France, 1978, and the Docteur ès-Sciences diploma from University Paul Sabatier, Toulouse, France in 1982. He joined LAAS in Toulouse in

1976, and since 1979, he has held a research position at the French Research Center (C.N.R.S.). His main fields of interest include system theory, control under constraints, optimization and stochastic systems.



Peter Lee is a Reader in the Computer Aided Process Engineering Group at the Department of Chemical Engineering, The University of Queensland. He has been conducting research in advanced process control for the past seven years. His primary research interests are in the use of model-based control algorithms, particularly using nonlinear process models. Prior to joining the

department, he worked at ICI Australia Engineering designing, installing and commissioning computer-based control systems. Dr Lee is a member of the Institution of Chemical Engineers (London).



Keum S. Hong was born in Taegu, Korea, in 1957. He received the B.S. degree in mechanical design and production engineering from Seoul National University, Korea, in 1979 and the M.S. degree in mechanical engineering from Columbia University in 1987. He is currently working for his M.S. degree in applied mathematics and Ph.D. degree in mechanical engineering

at the University of Illinois at Urbana-Champaign. From 1982 to 1985 he was a research engineer at Daewoo Heavy Industries, Ltd, Korea, solving vibrational and acoustic problems for diesel engines. His research interests are in nonlinear systems theory, control of distributed parameter systems, and adaptive control theory.



J. Lévine was born in 1950 in Paris, France. He obtained his Doctorat de 3ème cycle and his Doctorat d'État at the University of Paris-Dauphine in 1976, 1984, respectively. He is the head of the Centre Automatique et Systèmes (Center for Automatic Control and System Theory—previously Section Automatique, CAI) of École des Mines de Paris. His research interests include non-

linear control theory and its applications, nonlinear filtering and nonlinear identification.



Leonardo Lanari was born in Pesaro, Italy, on 10 January 1963. In 1987 he received the Laurea in electronic engineering and in 1989 he took the Masters degree in systems engineering from the Università di Roma "La Sapienza", where he is currently working towards the Ph.D. degree. His current research interests are in modeling and nonlinear control of flexible robotic structures.



Duan Li was born in Shanghai, China. He graduated from Fudan University, Shanghai, China, in 1977 and received the M.E. degree in control engineering from Shanghai Jiaotong University, Shanghai, China, in 1982, and the Ph.D. degree in systems engineering from Case Western Reserve University, Cleveland, Ohio, U.S.A., in 1987.

From 1977 to 1979 he was an Assistant Engineer in the Shanghai Institute of Instruments for Industrial Automation, Shanghai, China. From 1982 to 1983 he was on the faculty of Shanghai Jiaotong University. He has been a faculty member at the University of Virginia, Charlottesville, Virginia, U.S.A., since 1987. He is currently a Research Assistant Professor in the Department of Systems Engineering. His research interests are large-scale multiobjective systems theory, dynamic programming and multiobjective control. He has authored and co-authored over 30 technical papers. He was the guest co-editor of the special issue on multiobjective discrete dynamic systems in *Control—Theory and Advanced Technology*.



Jay H. Lee was born in Seoul, Korea on 10 March 1965. He received a B.S. in chemical engineering from the University of Washington, Seattle in June 1986 and a Ph.D. from Caltech, CA, in 1991, also in chemical engineering. Currently, he is an Assistant Professor in the Department of Chemical Engineering at Auburn University, AL. He is a member of Tau Beta

Pi Honor Society. His research interests include robust inferential control, analysis and control of multi-rate sampled-data systems, control structure selection and decentralized control.



Peter McIntosh is the Engineering and Technology Manager at Queensland Alumina Limited, Gladstone, Queensland, Australia. His department's responsibilities include raw material and energy efficiencies and process control computer applications within the plant.



Peter M. Mills is presently the Senior Research Engineer (Process Control) with CRA Advanced Technical Development in Australia. He received the B.E. degree from the University College of Central Queensland in 1980 and the M.E. (Research) degree from the University of Queensland in 1989. From 1980 to 1988 he worked at Queensland Alumina, applying conventional

and advanced control schemes in distributed control system environments.

He is presently working toward the PhD degree at the University of Western Australia. His current research interests are in nonlinear process control using neural networks.



Manfred Morari is a Professor and Executive Officer for Chemical Engineering at the California Institute of Technology. He received his diploma in chemical engineering from the Swiss Federal Institute of Technology in 1974 and the Ph.D. degree from the University of Minnesota in 1977. He was on the faculty of the University of Wisconsin from 1977 until 1983, when he assumed

his current position. He also held short-term positions with Exxon Research and Engineering, and ICI. His interests are in the area of process control and design, in particular robust and decentralized control and the effect of design on operability. He is the principal author of the book *Robust Process Control* published by Prentice-Hall (1989). Professor Morari has received several academic honors including the Donald P. Eckman Award of the American Automatic Control Council in 1980; the Allan P. Colburn Award of the AIChE in 1984; and the NSF Presidential Young Investigator Award also in 1984.



Edoardo Mosca received the Dr Eng degree from the University of Rome, Italy, in 1963 and the Libera Docenza in 1971. From 1964 to 1968 he was on the technical staff of Selenia S.p.A., Rome, Italy. From 1968 to 1972 he was associated with the Institute of Science and Technology, University of Michigan, Ann Arbor, U.S.A. During the academic year 1971-1972 he was

Visiting Associate Professor in the Department of Electrical Engineering, McMaster University, Hamilton, Ontario, Canada, and in the Department of Electrical Engineering, University of Naples, Naples, Italy. From 1972 to 1975, he was Associate Professor at the University of Florence, Florence, Italy, where, since 1975, he has been Professor at the chair of System Theory and, from 1980 to 1987 Chairman of the Dipartimento di Sistemi e Informatica. From 1984 to 1987 he was responsible for coordinating a national research project on system and control engineering. He has worked on statistical theory of signal detection and processing in radio communication and radar, representation of stochastic processes, estimation, system identification and adaptive control. His current research interests are generally in robust control and adaptive filtering and control. Dr Mosca is a

member of the IEEE; the Scientific Council of the Italian Group of Automatic and System Engineering of the National Research Council (CNR); the IFAC Technical Committee on Theory; and an editor of the *International Journal of Adaptive Control and Signal Processing*.



Paolo Nistri was born in Florence, Italy, in 1948. He received the Dr degree in Mathematics from the University of Florence in 1972. From 1974 to 1979, he was Assistant Professor at the University of Calabria, Cosenza, Italy. In 1980 he moved to the Engineering Faculty of the University of Florence, where, since 1982, he has been Associate Professor in the Systems and Informatics Department. His main research interests are: topological methods in nonlinear functional analysis, dynamical systems and mathematical control theory.



Yoshikazu Nishikawa was born in Shiga Prefecture, Japan on 18 March 1933. He received the B.E., M.E. and Ph.D. degrees in electrical engineering all from Kyoto University, in 1955, 1957 and 1962, respectively. Since 1960, he has been with the Department of Electrical Engineering at Kyoto University, where he is presently Professor in charge of the Instrumentation and

Control Laboratory. During the period September 1966 to March 1968, he was Visiting Assistant Professor of Engineering at UCLA, Los Angeles, CA, and was working at NASA in 1976. His present research interests include optimal/robust control, systems optimization, energy systems and distributed-autonomous systems including bio/neural systems. Dr Nishikawa is a member of several academic institutes, and currently he is serving as Chairman of the Working Group on Energy Systems Management and Economics, EMSCOM of IFAC.



Giuseppe Oriolo was born in Taranto, Italy, on 6 August 1962. In 1987 he received the Laurea in electronic engineering and in 1989 he took the Masters degree in systems engineering from the Università di Roma "La Sapienza", where he is currently working towards the Ph.D. degree. His current research interests are in trajectory planning and motion control of conventional and redundant robot arms.



Jonathan Partington obtained his Ph.D. in the Pure Mathematics department of the University of Cambridge in 1980 and continued to work in pure Functional Analysis until 1985, when he moved to the Cambridge University Engineering Department. Since 1989 he has been a lecturer in the School of Mathematics at the University of Leeds. He has written one book, an introduction

to Hankel operators, and his current research interests include applications of functional analysis to systems theory.

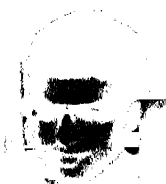


Nobuo Sannomiya was born in Wakayama City, Japan on 23 August 1939. He graduated from Kyoto University in Japan in 1962 and received the M.E. and Ph.D. degrees in electrical engineering from Kyoto University in 1964 and 1969, respectively. From 1967 to 1986, he was with the Department of Electrical Engineering, Kyoto University. Since 1986, he has been Professor of Electronics

and Information Science Department at the Kyoto Institute of Technology. His present research interests include optimal control theory, modeling and optimization techniques, and their applications.



Richard J. Perry received his Ph.D. in electrical engineering from Drexel University in 1981. Since then he has been Assistant Professor of Electrical Engineering at Villanova University, Pennsylvania. His research interests are in the areas of multivariable systems, computational algorithms, and VLSI array processors.



Jeff S. Shamma was born in New York, NY, in November 1963. He received the Ph.D. degree in 1988 from the Massachusetts Institute of Technology, Cambridge, MA. From June 1988 to August 1989, he was with the Laboratory for Information and Decision Systems at MIT, MA. Since September 1989, he has been an Assistant Professor of Electrical Engineering at the

University of Minnesota, Minneapolis, MN. His main research interest is robust control for nonlinear, time-varying, and adaptive systems.



P. Rouchon was born in 1960 in Saint-Etienne, France. After graduating from the Ecole Polytechnique in 1983, he obtained his Doctorat at Ecole des Mines de Paris in 1990. His research interests include chemical process modeling, simulation and control, dynamical system theory and nonlinear control theory.



Richard M. Stephan received the B.Sc. degree in electrical engineering from Instituto Militar de Engenharia (IME), Rio de Janeiro, in 1976, the M.Sc. degree from Universidade Federal do Rio de Janeiro (UFRJ/COPPE) in 1980, and the Dr.-Ing. degree from Ruhr-Universität Bochum, Germany, in 1985. During 1977 he worked as an engineer at Furnas Centrais

Elétricas, Rio de Janeiro. Since 1978 he has been at the Universidade Federal do Rio de Janeiro (UFRJ). From 1982 to 1985 he was on leave from UFRJ as a DAAD (Deutscher Akademischer Austauschdienst) scholar at Ruhr-Universität Bochum. His research interests are in the control of electrical drives, power electronics and power systems.



Berk Rustem was born in Istanbul, Turkey in 1946. He obtained a B.S. in mechanical engineering from Robert College, Istanbul, an M.Sc. in systems engineering and a Ph.D. in computer science from the University of London. From 1982 to 1989 he was SERC Advanced Fellow and Senior Lecturer and is currently Reader in the Department of Computing at Imperial College of Science,

Technology and Medicine, London. He is the author and editor of two books on nonlinear programming algorithms and decision making and control in economic systems. He is the editor of *Journal of Economic Dynamics and Control* and an associate editor of *Computer Science in Economics and Management*. He was the chairman of the international program committee of the 6th IFAC/SEDC/IFORS/IFIP Symposium on Dynamic Modelling and Control of National Economies, Edinburgh, 1989, and is currently chairman of the IFAC/EMSCOM working group on national and regional economies.



Hsu H. Sun received the B.S.E.E. degree from Chiao Tung University, Shanghai, China, in 1946, the M.S.E.E. degree from the University of Washington, Seattle, in 1949, and the Ph.D. degree from Cornell University, Ithaca, NY, in 1955.

He joined Drexel University, Philadelphia, PA, in 1953, and served as Director of Biomedical Engineering, from 1964 to 1974,

and Chairman of Electrical Engineering, from 1973 to 1978. He is now an Ernst O. Lange Professor of Electrical Engineering at Drexel University, and an Adjunct Professor of Physiology at Temple University, School of Medicine, Philadelphia, PA. His special interest is in the system engineering application to biomedical problems.

Dr Sun has served on various committees including the AdCom of the IEEE Engineering in Medicine and Biology Society, Translation Relations Committee, Standard Committee, and on the Board of Alliance of the Biomedical Engineering Societies. He has been editor-in-chief of the *IEEE Transactions on Biomedical Engineering*, member of Study Committee on Surgery and Bioengineering in NIH, member of Visiting Committee in Critical and Engineering Technologies Division in NSF and currently he is editor-in-chief of the *Annals of Biomedical Engineering* and on the Editorial Board of *Automatica* and a Fellow of the IEEE.



Panagiotis Tsiotras was born in Athens, Greece, on 20 September 1963. He received the Diploma in mechanical engineering from the National Technical University, Athens, Greece, in 1986, and the M.Sc. degree in aerospace engineering from Virginia Polytechnic Institute and State University, in 1987. During 1989 he was with the Interdisciplinary Center of Applied Mathematics at VPI &

SU as a research project assistant. Since August 1989, he has been with the School of Aeronautics and Astronautics at Purdue University, where he is currently pursuing his Ph.D. degree. He is a recipient of the David Ross Research Fellowship and the NATO Scholarship for doctoral research. His current research interests are in the areas of nonlinear systems analysis and control.



Takashi Tsuda received the B.S. in Physics from Osaka University in 1973 and joined Fuji Electric Corporation. In 1978 he was transferred to FUJI FACOM Corporation. Since then, he has been engaged in the research and development of software for process automation and factory automation. He is currently Manager of System Engineering Section No. 4, System Engineering

Department No. 6 of the Systems Engineering Group. He is a member of the Society of Instrument and Control Engineers.



Heinz Unbehauen received the Dipl.-Ing. degree in mechanical engineering from the University of Stuttgart, Germany, in 1961 and after additional studies in electrical engineering, whilst being employed as a research assistant, the Dr.-Ing. and Dr.-Ing. habil. degrees in control engineering from the same university in 1964 and 1969, respectively. In 1969 he was awarded

with the title of Docent and in 1972 he was appointed Professor in Control Engineering in the Department of Energy Systems at the University of Stuttgart. Since 1975 he has been Professor at Ruhr-University Bochum, Faculty of Electrical Engineering, where he is head of the Control Engineering Laboratory. He was a visiting Professor at Hokkaido University (Japan) in 1974, at IIT Madras (India) in 1975, Tianjin University (China) in 1983, Tongji University Shanghai (China) in 1986 and IIT Kharagpur (India) in 1987. He has authored and co-authored over 190 journal articles and conference papers and 7 books. His main research interests are in the fields of optimization, system identification, digital control, adaptive control, process control of multivariable and large scale systems. Dr. Unbehauen is Associate Editor of *Automatica* and *C-T&T* and serves on the editorial boards of the *International Journal of Adaptive Control and Signal Processing* and *Optimal Control Applications and Methods*, and is Editor of several Conference Proceedings.



Oded Yaniv was born in Israel in 1950. He received the B.Sc. degree (mathematics and physics) from the Hebrew University, Jerusalem, in 1974, and M.Sc. (physics) and Ph.D. (applied mathematics) degrees from the Weizmann Institute of Science, Rehovot, Israel in 1976 and 1984, respectively. His industrial experience includes developing microelectronics, 1976-1978, being

a development engineer at the Israel Aircraft Industries 1979-1980, and serving as a Senior Control Engineer at Tadiran (System Division), Holon, Israel 1983-1987. Since 1987 he has been a Lecturer at the Tel-Aviv University, Israel. His main research interests include synthesis of uncertain linear and nonlinear multi-input-output feedback systems.

